

UNCLASSIFIED

Copy 10 of 43 copies

AD-A286 831



IDA PAPER P-2949

SUMMARY REPORT OF
DEFENSE SCIENCE STUDY GROUP III
1992 – 1993

Volume I

N. P. Licato



December 1994

95-01663

Prepared for
Advanced Research Projects Agency

Approved for public release, distribution unlimited; 24 May 1995.

85 € 27 001 DTIC QUALITY INSPECTED 8



INSTITUTE FOR DEFENSE ANALYSES
1801 N. Beauregard Street, Alexandria, Virginia 22311-1772

UNCLASSIFIED

IDA Log No. HQ 94-45733

DEFINITIONS

IDA publishes the following documents to report the results of its work.

Reports

Reports are the most authoritative and most carefully considered products IDA publishes. They normally embody results of major projects which (a) have a direct bearing on decisions affecting major programs, (b) address issues of significant concern to the Executive Branch, the Congress and/or the public, or (c) address issues that have significant economic implications. IDA Reports are reviewed by outside panels of experts to ensure their high quality and relevance to the problems studied, and they are released by the President of IDA.

Group Reports

Group Reports record the findings and results of IDA established working groups and panels composed of senior individuals addressing major issues which otherwise would be the subject of an IDA Report. IDA Group Reports are reviewed by the senior individuals responsible for the project and others as selected by IDA to ensure their high quality and relevance to the problems studied, and are released by the President of IDA.

Papers

Papers, also authoritative and carefully considered products of IDA, address studies that are narrower in scope than those covered in Reports. IDA Papers are reviewed to ensure that they meet the high standards expected of refereed papers in professional journals or formal Agency reports.

Documents

IDA Documents are used for the convenience of the sponsors or the analysts (a) to record substantive work done in quick reaction studies, (b) to record the proceedings of conferences and meetings, (c) to make available preliminary and tentative results of analyses, (d) to record data developed in the course of an investigation, or (e) to forward information that is essentially unanalyzed and unevaluated. The review of IDA Documents is suited to their content and intended use.

The work reported in this document was conducted under contract DABW01 84 C 8054 for the Department of Defense. The publication of this IDA document does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that Agency.

UNCLASSIFIED

IDA PAPER P-2949

**SUMMARY REPORT OF
DEFENSE SCIENCE STUDY GROUP III
1992 – 1993**

Volume I

N. P. Licato

December 1994

Approved for public release, distribution unlimited; 24 May 1995.



INSTITUTE FOR DEFENSE ANALYSES

Contract DASW01 94 C 0054

ARPA Assignment A-103

UNCLASSIFIED

CONTENTS

Volume I

I.	INTRODUCTION.....	I-1
A.	Goal.....	I-1
B.	Program Characteristics.....	I-1
C.	Members	I-2
D.	Mentors and Advisors	I-2
E.	Alumni.....	I-2
F.	Sponsor	I-2
II.	THE PROGRAM.....	II-1
A.	Activities During 1992.....	II-1
1.	Session 1: February 12-14, 1992	II-2
2.	Session 2: June 22-30, 1992	II-2
3.	Session 3: August 10-14, 1992	II-4
4.	Session 4: November 16-18, 1992	II-5
B.	Activities During 1993.....	II-7
1.	Session 5: March 10-12, 1993	II-7
2.	Session 6: June 21-29, 1993	II-7
3.	Session 7: August 5-12, 1993.....	II-8
4.	Session 8: November 14-16, 1993	II-8
III.	CONCLUSION.....	III-1
A.	Program Continuation	III-1
B.	Alumni Activity.....	III-1
C.	ACDA Symposium	III-2
D.	DSSG IV, 1994-1995.....	III-2
IV.	STUDIES AND ANALYSES	
A.	Fatigue Monitoring of Critical Aircraft Components Using Multiple Microsensors.....	IV-1
B.	An Assessment of the State-of-the-Art in Constitutive Modeling and Hydrocodes for Ballistic Impact and Blast Waves.....	IV-25
C.	Technical Methods for Reducing Ground Forces Fratricide	IV-117
	Annexes:	
A.	Coded Mesochronous Coherent Detection	IV-145
B.	IFF with the Spatially Encoded Bar Code (SpEcBar).....	IV-151
D.	Some Comments on Autonomous Target Recognition (ATR) Research.....	IV-159

E. Lateral Wave Modifications for Electromagnetic Propagation Near Interfaces	IV-177
Annexes:	
A. Explicit Sample of a Sommerfeld Integral Representation.....	IV-193
B. Ordinary Dipole Radiation Components in Free Space	IV-197
C. Sample Wu-King Dipole Solution Components.....	IV-201
F. Conflict and Interest: Siting a Nuclear Waste Repository	IV-211
G. New Detector System for Airport Security Against Bombing Threats.....	IV-243
H. Tracing the Origin of Mycotoxin CBW Agents by ^{13}C Isotopic Fractionation.....	IV-265
I. Basic Skills Training for the Civilian Workforce: What Can Be Learned From the U.S. Military?	IV-279
J. Benefits of Increased Automation for Nuclear Naval Reactor Operation and Personnel Training.....	IV-303

Appendices

A. Members	
B. Mentors and Advisors	
C. DSSG Management Team	
D. Program Plan	
E. Alumni	
F. Some Defense Related Activities of DSSG Alumni	
G. Glossary	
H. Distribution List for IDA Paper P-2949, Volume I	

Volume II
(published separately)

V. CLASSIFIED STUDY AND ANALYSIS

PIRANHA: A Precisely Launched Charges Concept For All-Sector Surface Ship Torpedo Defense	V-1
Annex:	
A. Fortran Source Code for Piranha Simulation	V-29

Appendix

A. Distribution List for IDA Paper P-2949, Volume II	
--	--

I. INTRODUCTION

This report summarizes the activities of the Defense Science Study Group (DSSG) from 1992 through 1993. The DSSG is a program of education and study directed by the Institute for Defense Analyses (IDA) and sponsored by the Advanced Research Projects Agency (ARPA). Members of the DSSG are generally young science, engineering, and mathematics professors, who have achieved national recognition in their fields and who are among the likely future leaders of science and technology.

A. GOAL

The program's goal is to foster a long-term interest in national security issues among DSSG members, an interest that would lead to service in advisory roles and a continuing involvement in research related to such issues. The program is intended to convey to the members an understanding of the technical complexities of national security issues and an appreciation for the people and operations involved. The program also solicits new insights from the members and helps facilitate their continued involvement with problems of national security.

B. PROGRAM CHARACTERISTICS

The DSSG program is concerned with education in the broad field of national security. For approximately 20 days each year over a 2-year period, members have the opportunity to focus on defense policy, related research and development, and the systems, missions, and operations of the military services.

An appreciation of the defense environment is gained by attending presentations made by technical specialists and senior officials of the Department of Defense and the defense industry. Members visit various military bases throughout the United States to see the operating forces and to meet with senior commanders and other members of the military services. Visits are also made to defense laboratories and industrial facilities to gain a perspective on design and manufacturing technology for current and future systems. In addition, members of the DSSG attend presentations made by senior members of organizations such as the Department of Energy (DoE), the Arms Control and Disarmament Agency, (ACDA) the Central Intelligence Agency (CIA), and by Congressional staffs. Site visits are also made to some of these organizations.

As part of the education program, the DSSG members, individually or in small groups, also conduct "initial inquiries" or prepare "think pieces" on national security problems of their choice. A high degree of flexibility exists in the choice of problems and the manner in which they are studied.

C. MEMBERS

IDA solicits nominations from major universities; from mentors, advisors, alumni, and current members; and from Government organizations such as ARPA. In addition, candidates are identified from among Sloan Research Fellows and those who have received awards from organizations such as the National Science Foundation. The nominees are generally faculty members between 30 and 40 years of age. Since a security clearance is required, members must be U.S. citizens. After an extensive process of evaluating qualifications and checking references, IDA invites approximately 15 candidates to become DSSG members.

D. MENTORS AND ADVISORS

A group of mentors and advisors who have distinguished careers in defense, industry, or academia are closely associated with the DSSG. They help guide IDA and ARPA on the conduct of the program, suggest study topics, counsel and work with members during their studies, and help provide access to places and organizations involved in national security.

E. ALUMNI

The DSSG program is an investment in the future. IDA and ARPA make every effort to ensure that former DSSG members are offered opportunities for continued involvement in areas of national security. Such opportunities include serving as advisors, consultants, or members of panels, study groups, and task forces for organizations that are addressing technological problems of national importance.

To assist former DSSG members in keeping abreast of key issues of national concern, all alumni are invited to return to IDA periodically to receive briefings on current problems of national security, to exchange information with fellow alumni, and to meet the new members of the DSSG.

F. SPONSOR

As the program sponsor, ARPA provides overall guidance to the program, assists in developing the program's technical agenda, and directly benefits from the results of the DSSG's activities.

II. THE PROGRAM

During the early part of FY 1992, 14 new members were selected to the DSSG for a 2-year period. A list of those members and a summary of their background can be found in Appendix A.

A few adjustments were made to the mentor portion of the program. All current mentors were asked to serve for an additional 2 years. New mentors will serve a 4-year term subject to 2-year extensions as agreed on by each mentor and IDA. Two new mentors were added this year: Steven Koonin, an alumnus of the first DSSG class and Professor of Physics at the California Institute of Technology; and Curtis Callan, Professor of Physics at Princeton University and current Chairman of JASON. Of special importance is the close association that some mentors will have with individual members as they work on study projects. Also, a new role of advisor has been created for those who are willing to assist on occasion and promote the program but who generally do not attend meetings. Appendix B lists the mentors and advisors during this period.

The program is managed at IDA by the DSSG Management Team, shown in Appendix C. A detailed 2-year program plan is presented to members, mentors, advisors, ARPA, and others to ensure that all involved in the program will know well in advance approximately what is scheduled for the 2-year period. However, the program plan is flexible to accommodate the changing times, the guidance of the mentors, special interests of the members, and special requests of ARPA. Appendix D contains the 2-year program for the DSSG.

A. ACTIVITIES DURING 1992

During the first year of this 2-year program, the DSSG members were introduced to the Department of Defense, the military services, other major Government agencies (such as the Department of Energy), and a variety of technical problems of national interest. The DSSG sessions during the first year are summarized in Table 1 and discussed in more detail in the following.

Table 1. DSSG Schedule for 1992

Session	Activity	Dates
1	Introduction	February 12-14
2	Tour of Air Force and Aerospace Manufacturing Facilities	June 22-30
3	Visit to Army and Navy Facilities	August 10-14
4	Briefings on Congress, Government Agencies, and Review of Study Topics	November 16-18

1. Session 1: February 12-14, 1992

At the first meeting new members were introduced to the DoD, the military services, and IDA. Specifically, Dr. Victor Reis, Director, Defense Research and Engineering (DDR&E), provided information on the major science and technology thrust areas; Dr. Gary Denman, Director of ARPA, gave an overview of ARPA; Mr. Philip Major, Vice President-Planning and Evaluation at IDA, provided information on the organization of the DoD; Dr. Robert Roberts, Vice President-Research at IDA, briefed the members on IDA, its divisions, and customers within the DoD. Technological aspects of major issues and problems within the Army, Navy, and Air Force, were presented by MG Jerry Harrison, USA, Commanding General, U.S. Army Laboratory Command; Dr. Fred Saalfeld, Director, Office of Naval Research; and MajGen Robert Rankine, USAF, Deputy Chief of Staff, Technology; respectively. Dr. Steven Koonin briefed the members about the DSSG from the perspective of an alumnus and Dr. Richard Garwin, IBM Fellow and Science Advisor to Director-Research, IBM, spoke to the group about his long association with the Government as a consultant and advisor.

During this session, the group visited the new IDA Simulation Center; the National Military Command Center in the Pentagon, where they were briefed by ADM David Jeremiah, Vice Chairman of the Joint Chiefs of Staff; and the U.S. Army Night Vision & Electro-Optics Directorate at Ft. Belvoir, VA, where they received a tour of the laboratory and the opportunity to use night vision devices. The session concluded after a round table discussion at IDA with suggestions by mentors and discussions by members of relevant study topics or areas for consideration by the members during the second year.

2. Session 2: June 22-30, 1992

The DSSG spent this session visiting aerospace manufacturing and Air Force facilities. The tour began with a visit to TRW where the DSSG was briefed on the Military Strategic-

Tactical and Relay (MILSTAR) satellite system, the Defense Support Program (DSP), Brilliant Pebbles, and Brilliant Eyes. The group then visited Rockwell for a review of the B-1B aircraft program and the AC-130U gunship; Northrop for the B-2 aircraft program; Lockheed for the F-117 aircraft and stealth materials programs, and Martin Marietta to learn about intelligent systems, Titan production, and advanced materials.

Air Force visits included Nellis AFB, NV, to observe the Red Flag training exercise; Buckley Air National Guard Base for satellite communications; the North American Aerospace Defense Command (NORAD) for an introduction to air defense, space surveillance, and missile warning; and finally the newly formed Strategic Command to learn about their missions. During the tour, two in-flight refueling exercises were scheduled. On the first flight from a C-141, the members observed an air-to-air refueling by a KC-135. On the second flight, each member had an opportunity to operate the refueling boom on a KC-135.

At the conclusion of this session, the members listed ideas for possible study. Table 2 lists these topics.

Table 2. Study Topics Considered During Session 2

Sensors (MCT), raised performance/lower weight
EW electronic countermeasures
Personal & vehicular IFF
Integrated aircraft network
Methods for integrating commercial technology into military
Automated mission planning, especially for aircraft
Optical (visible) stealth
Autonomous aircraft (recon & strike)
Minimum effective industrial capability
Electronic spotlight capability
Lockdown, satellite, over the horizon (OTH) radar, & stealth technology
IR stealth
Feature detection in signals from radar, infrared (IR), etc.
Robust, portable, field, high bandwidth receiver
Effect of stealth on nuclear deterrence
Conventional deterrence
Water absorption in composites
Focal plane array on IR missiles & countermeasures
Efficacy of large forces on worldwide basis
Autonomous target recognition
Nuclear weapon decommissioning
Cruise missile detection
Effect of proliferation of stealth technology
Conversion of industrial military capability to civilian uses
Effect of military on biodiversity

Table 2. (continued)

Role of human in the loop (make easier & elimination)
Classification vs. rate of technology development
Military procurement; tracking from initial investment (patterns of distribution of contracts)
Does military education have something to offer civilian education?
Remote detection of chemical & biological weapons (battlefield & factories)
Influence processing and materials on shaped charges
Ceramic composites on aircraft
Shaped memory alloys for active radar cancellation
Concurrent engineering & manufacturing
Distributed vs. centralized threat detection and processing
Sensor blinding & protection
Brain wave amplification for communication & man-machine interface -- NMR Pet-Scan applications
Biologically based security systems
Location of a single (particular) individual (uncooperative)
Direction & effect of new technology on future arms races
Late targeting of strike aircraft
Simulation of massive exercises
Defensive vs. offensive weapons systems
New age cannon technology

3. Session 3: August 10-14, 1992

The DSSG visited Navy and Army facilities during this session. In Norfolk, VA, the DSSG heard the morning briefing presented to the Commander-in-Chief Atlantic Fleet and met with him for questions and answers; viewed a display of Navy aircraft, including the F-14 and A-6, and talked to pilots and crew members; toured a Landing Craft Air Cushion (LCAC) vehicle and saw a demonstration of its performance over water and land. In addition, the members traveled by helicopter to the destroyer, U.S.S. *SCOTT* (DDG995), at sea. They were given a tour of the ship, talked with the crew and had lunch with the captain and officers. The group also visited the submarine base in Kings Bay, GA, where they toured the training facility, strategic weapons facility, the missile assembly building, and the Trident submarine, U.S.S. *WEST VIRGINIA*.

The DSSG tour ended at Ft. Bragg, NC, where the members visited the U.S. Army XVIII Airborne Corp, the 82nd Airborne Division, and the Special Operations Command. The group received briefings and modular demonstrations by the Army Special Operations Command (Airborne); toured areas for parachute rigging, preparation of heavy equipment for drop, and observed a ground simulation of jump procedures. A special reception hosted by the Commanding General was held at which the group had the opportunity to

meet many of the senior officers. For one-half day an enlisted person was assigned as a "buddy" to each attendee. A "buddy" accompanied each member through the static display of equipment that is dropped by parachute and helped each member put on a parachute. They later joined the members during a field dinner in a tent. The DSSG felt that this was a very effective way for them to get to know these young people and some of their responsibilities.

At the conclusion of this session, the members listed ideas for possible study. Table 3 lists these topics.

**Table 3. Study Topics
Considered During Session 3**

Active/reserve force mix
Competing approaches across Services
Critique of basic force plan (force projection)
Nuclear proliferation/war plan
Acoustical stealth/active noise cancellation
Bar code inventory, tracking, identification
Light vs. heavy transport
Torpedo defense; e.g., last ditch defense
Hydrofoils
Lighter rifles
O ₂ generation without H ₂ dumping
Frequency smearing on submarines
Sonar-based autonomous recognition system on submarines
Distributed simulation

4. Session 4: November 16-18, 1992

At the end of Session 3, the DSSG members were asked to choose one or more candidate topics for individual or team study and send a short description to IDA. During the first day of this session, the members and mentors discussed these study topics to help decide on the feasibility of the topic, to suggest ways to proceed, and to identify sources of information. Mentors were encouraged to select a study topic where he/she can best assist the member or team. Table 4 contains the tentative study topics.

The second day consisted of an introduction to Congress and its role in national security and to the Executive Branch's Office of Management and Budget (OMB). The group heard briefings on the House and Senate Armed Services Committees, Office of Technology Assessment (OTA), and OMB from representatives of these organizations. In addition, Dr. John Deutch of MIT spoke to the group on science, technology, and national security, and Lieutenant General Brent Scowcroft, USAF (Ret.), Assistant to the President

for National Security Affairs, spoke to the group on national security. On the last day, speakers from other major Government agencies spoke to the DSSG; they included Dr. David Chu, Assistant Secretary of Defense (Program Analysis and Evaluation), on long-term planning for defense; Dr. Everett Beckner from DoE on national defense from the perspective of DoE; Mr. Paul Wolfowitz, Under Secretary of Defense (Policy), on America's role in the new security environment; and Ambassador Linton Brooks of Arms Control and Disarmament Agency on nuclear arms control after the cold war. Also, IDA presented a current IDA research topic, monopulse tracking of ground targets, and an overview of the Supercomputing Research Center, an IDA division located in Maryland.

Table 4. Tentative Study Topics for 1993

Detection of chemical warfare agents and explosives
Coded mesochronous coherent detection for IFF
Chemical and biological sensing
Methods for integrating modern industrial technology into military (especially Army) operations
Minimum effective industrial levels required for short-time reconstitutable military reconversion
Bar-code technology applications for military use
A study of effective warfighting strategies within the confines [of] multi-party coalitions
Defining strategies in wars of compellance
Analysis of FEWS detector systems
Nuclear weapons decommissioning and waste management
Redefining national security in the post cold war world
Infrared stealth
Acoustical detection in shallow water
Base force and force projection strategies
Autonomous target recognition
Computing architecture
IR imaging to detect aircraft
Protection of optical sensors from laser blinding
Improving description of material anisotropy and damage evolution for analysis of critical components
A thermomechanical technique for submarine detection under arctic ice cover
What can the American civilian educational system learn from the military?
Alternative methods for oxygen generation on submarines
Conversion of decommissioned military bases to civilian use
Use of naval reactor simulators for training in the nuclear navy
Detection of cruise missiles and stealth aircraft by photon echo probing
Gas management on submarines
Last-second torpedo defenses
Bioluminescence for submarine detection
Uncooled thermal-based weapons
Fatigue monitoring of critical aircraft components using multiple microsensors
High temperature integrated microsensors
Lightweight structural materials for infantry applications
Welding of titanium alloys
Processing and applications of monolithic NiAl alloys

B. ACTIVITIES DURING 1993

The DSSG sessions during the second year are shown in Table 5. The objectives for this year were to assist the members in finalizing topics for study, provide time and facilities for the members to conduct and complete their study topics, and provide an introduction to the intelligence agencies.

Table 5. DSSG Schedule for 1993

Session	Activity	Dates
5	Visits to Intelligence Agencies	March 10-12
6	Visits to Manufacturing and Advanced Technology Facilities and Study Session at IDA	June 21-29
7	Study Session at LANL and Visit to NTC	August 5-14
8	Final Session and Presentation of Study Topics	November 14-16

1. Session 5: March 10-12, 1993

During this session, the DSSG visited the Defense Intelligence Agency (DIA), the Central Intelligence Agency (CIA), and the National Security Agency (NSA), where they were briefed on the agencies and their activities and also received briefings by the Directors of DIA, CIA, and NSA: LtGen James Clapper, USAF; Mr. R. James Woolsey; and VADM J.M. McConnell, USN, respectively.

2. Session 6: June 21-29, 1993

This session began with visits to Raytheon Co., the MITRE Corp., and MIT Lincoln Laboratory, all located in the Boston area. At Raytheon, briefings were presented on ground-based radar and associated systems and on naval defense systems and commercial variants. In addition, the group visited the Missile Systems Division Plant and saw Patriot missiles in production.

At MITRE, briefings were presented on software acquisition and on an education initiative, in addition to demonstrations and briefings on the Imagery Laboratory, the Virtual Reality Laboratory, mission planning, and the Joint Surveillance, Target, Attack Radar System (JSTARS) program at Hanscom AFB. The visit to MITRE provided more information to the DSSG about how Federally Funded Research Development Centers (FFRDCs) operate and their missions.

At Lincoln Laboratory, the group received briefings on automatic target recognition, air vehicle survivability evaluation, space-based visible surveillance, theater missile

defense, high temperature superconductive microwave components and systems, and tapered semiconductive-diode optical power amplifier. They also toured the Optical Communications Laboratory.

The group returned to IDA and from June 24-29 decided on their study topics, began their research, and met with knowledgeable individuals within and outside IDA to discuss their ideas.

3. Session 7: August 5-12, 1993

The DSSG continued their studies at the Los Alamos National Laboratory and while there toured some of their labs and facilities. After leaving the laboratory, the DSSG traveled to Ft. Irwin, CA, to visit the National Training Center and see unrehearsed "warfare" of U.S. infantry and armor brigades against the NTC's Opposing Force.

4. Session 8: November 14-16, 1993

Members presented results of their study topics to the mentors, ARPA, and IDA. Table 6 lists the DSSG researchers and their studies. The study reports are included in Section IV with the exception of the study topic, PIRANHA, which is included in Volume II. This was the completion of the 2-year program of DSSG 1992-1993.

Table 6. Publications of DSSG III

DSSG Member	Study Topic
Dennis L. Polla	Fatigue Monitoring of Critical Aircraft Components Using Multiple Microsensors
David L. McDowell	An Assessment of the State-of-the-Art in Constitutive Modeling and Hydrocodes
Peter W. Voorhees	for Ballistic Impact and Blast Waves
William J. Dally	Technical Methods for Reducing Ground Forces Fratricide
Kevin K. Lehmann	Annexes: A. IFF with the Spatially-Encoded Bar Code (SpEcBar) B. Coded Merochronous Coherent Detection
Robert A. Hummel	Some Comments on Autonomous Target Recognition (ATR) Research
Robert A. Hummel	Lateral Wave Modifications for Electromagnetic Propagation Near Interfaces
S. James Gates, Jr.	Conflict and Interest: Siting a Nuclear Waste Repository
Nancy M. Haagel	New Detector System for Airport Security Against Bombing Threats
Peter Chen	Tracing the Origin of Mycotoxin CBW Agents by ¹³ C Isotopic Fractionation
Mark E. Davis	
Peter Chen	Basic Skills Training for the Civilian Workforce: What Can Be Learned From the U.S. Military?
Anne B. Myers	
Gerald A. Navratil	Benefits of Increased Automation for Nuclear Naval Reactor Operation and Personnel Training
Thomas C. Halsey	PIRANHA: A Precisely-Launched-Charges Concept For All-Sector Surface Ship
Robert A. Pascal, Jr.	Torpedo Defense

III. CONCLUSION

A. PROGRAM CONTINUATION

At the request of Dr. Gary Denman, Director of ARPA, three DSSG members (William Dally, David McDowell, and Gerald Navratil) met with him on the morning of January 4, 1993, to discuss their views of the program and its importance to them. On January 21, General W.Y. Smith and Dr. Julian Nall met with Dr. Denman to discuss the next group for 1994 and 1995. Dr. Denman verbally approved the next 2-year program, and in March IDA began the search for candidates.

B. ALUMNI ACTIVITY

During this period there were 31 alumni of the DSSG. They continue to play a role in national security and to be of assistance to ARPA. Also, research of interest for the DoD is being carried out by various alumni. Four DSSG alumni have joined JASON, one before he completed the DSSG program. One alumnus is a member of the Defense Science Board.

As an example of one of the many studies that alumni Steven E. Koonin and Nathan S. Lewis have been involved in as members of JASON is a 1993 JASON summer study sponsored by ARPA on a bio-medical technology initiative related to health care.

During the previous DSSG class, two alumni, Werner J.A. Dahm and Daniel M. Nosenchuck developed a new concept for the control of tactical weapons. Upon completion of DSSG II, the authors continued their study under IDA sponsorship, expanding the concept to strategic weapon systems. The report on this study is IDA Paper P-2783, *AS-IS (Active Safing and Isolation System): A Satellite-Based Continuing Authorization Concept with Application to Control of Naval Strategic Nuclear Missiles and Tactical Weapons*, June 1992.

For another alumnus, Frederick K. Lamb, an interest in under ground nuclear testing began with his participation in the first DSSG class. He has maintained a continuing interest in this area, and the Nuclear Monitoring Office of ARPA (through IDA) sponsored a study of technologies for verifying limits on underground nuclear testing with Professor Lamb as the researcher. IDA Document D-1363, *Effects of Nuclear Devices and Device Canisters on the Accuracy of Hydrodynamic Yield Estimates for Threshold Test*

Ban Treaty and Peaceful Nuclear Explosions Treaty Verification, describes some of the results that have emerged from Professor Lamb's study of technologies for onsite verification of nuclear test bans.

Appendix E lists the alumni of the DSSG and Appendix F contains information on some of the defense-related activities of the alumni.

C. ACDA SYMPOSIUM

A valuable outgrowth of the DSSG program was a 2-day symposium sponsored by ACDA, which was held February 4-5, 1993. The symposium featured presentations by senior members of the arms control community. The program was designed to introduce the DSSG alumni to arms control policy, research and development (R&D), systems, missions, and operations. By sponsoring the symposium, ACDA hoped to engage this group of leading young U.S. academics in issues germane to arms control and to establish a close relationship between them and ACDA. A summary of the symposium is reported in IDA Document D-1351, *Arms Control and Disarmament Agency Symposium, 4-5 February 1993, Summary of Proceedings*, April 1993.

D. DSSG IV, 1994-1995

An intensive search and referral process was conducted during 1993 for members of the next DSSG program. Letters were sent to 61 universities throughout the country requesting that each nominate one or two of their best young faculty members. Nominations were also requested from DSSG alumni, current members, mentors, JASON members, the Chief of Naval Research, the new DDR&E, and others. By mid-July, the new members were chosen. In addition to academic excellence, consideration was also given to discipline and geographic distribution.

IV. STUDIES AND ANALYSES

**A. FATIGUE MONITORING OF CRITICAL AIRCRAFT
COMPONENTS USING MULTIPLE MICROSENSORS**

**Dennis L. Polla
University of Minnesota
Minneapolis, Minnesota**

ABSTRACT

Microsensors for crack detection and fatigue monitoring in critical aircraft components is proposed as a means of improving the reliability and useful life of defense aircraft. Microelectromechanical systems (MEMS) technology based on well-developed extensions of integrated circuit manufacturing methods is proposed as an enabling technology to form multiple, inexpensive, small size, and low weight material monitoring devices.

Two technical approaches are discussed. One approach is based on high-frequency acoustic emission monitoring for both real-time and ground-based servicing. A second approach is based on the use of integrating fatigue strain gauges. Both sensors are compatible with the fabrication of on-chip signal conditioning electronics.

Some requirements, advantages, and limitations of these material sensing approaches are discussed. Preliminary feasibility experiments are presented.

FATIGUE MONITORING OF CRITICAL AIRCRAFT COMPONENTS USING MULTIPLE MICROSENSORS

1. PROBLEM STATEMENT

Several U.S. military aircraft and helicopter models are nearing the end of their original design lifetimes. Failures, sometimes with injury and even loss of life, are increasing and generally associated with high use. Current methods available to reduce the rate of failures are limited and often difficult to apply. A significant fraction of helicopter failures occur in transmission components and mechanical linkages between the engine and the rotor blades. Aircraft failures often occur in highly loaded areas such as areas near the wing attachment lugs and fastener holes. The failures occur in the form of cracks usually initiated at locations of high stress. Microcracks continually form often growing quickly during flight to catastrophic size with little or no warning.

Replacement of the older models with new models, the most straightforward step, is unlikely in an era of limited budgets and contraction in the size of the defense establishment. Replacing critical components which are prone to the high failure rates with new components is being pursued, but this approach is also expensive and requires long lead times. The only other presently available option is aggressive maintenance which is expensive, time-consuming, and not 100% effective since many fatigue susceptible parts of an aircraft are inaccessible without extensive disassembly. The general perception among aviators is that maintenance procedures as a way of reducing failure rates, has reached a plateau. In many instances two hours are spent on maintenance for every hour of flight time.

This paper presents two approaches toward minimizing the occurrence of aircraft failure using newly developed microelectromechanical systems (MEMS) sensor technologies [1-5].

The first approach makes use of multiple embedded or surface-mounted acoustic emission sensors to detect material crack formation. This approach is based on the assumption that direct detection of the formation and growth of cracks is feasible through the use of sensors directly mounted on the components thought to be most likely to suffer from crack formation. This sensor continuously and directly monitors the component for the development of microcracks which ultimately cause component failure. The acoustic

emission sensor and associated signal conditioning electronics would produce an alarm to be given when significant crack growth is detected. The placement of multiple acoustic emission sensors on aircraft components could also allow triangulation methods to be used to identify crack location. The alarm would be given before complete failure has occurred and hopefully would give the pilot sufficient time to land safely.

Alternately, integration of the acoustic emission generated over time provides a quantifying method to determine aggregate crack growth. This information could be read out during ground-based servicing. Such readout schemes would be useful in determining the need for component replacement and could represent a cost savings through the elimination of unnecessary component replacement.

The second approach makes use of embedded micro-strain gauges to monitor the integrated fatigue associated with materials and components as natural flexing of components and joints takes place over time. The embedded strain gauges can be periodically interrogated during routine servicing to determine the amount of bending and strains over the last flight and record the number of flexural bending cycles experienced. This ground-based information can then be compared with failure models to determine whether repair or component replacement is necessary.

Both microsensor approaches described above are implemented using well-established silicon microsensor technologies. The inherent advantages of these microsensors and manufacturing methods include:

Table 1. Advantages of Microsensors Applied to Aircraft Fatigue Monitoring and Crack Detection

1. Low Unit Cost	Costs should be less than \$1/unit in batch quantity
2. Low System Weight	Typical weight of each sensor package is less than 10 gm
3. Multiple Information	Sensors can be distributed in multiple locations to directly gather information
4. Redundancy	Several microsensors can be placed side-by-side for both corroboration and replacement

2. TECHNICAL APPROACH

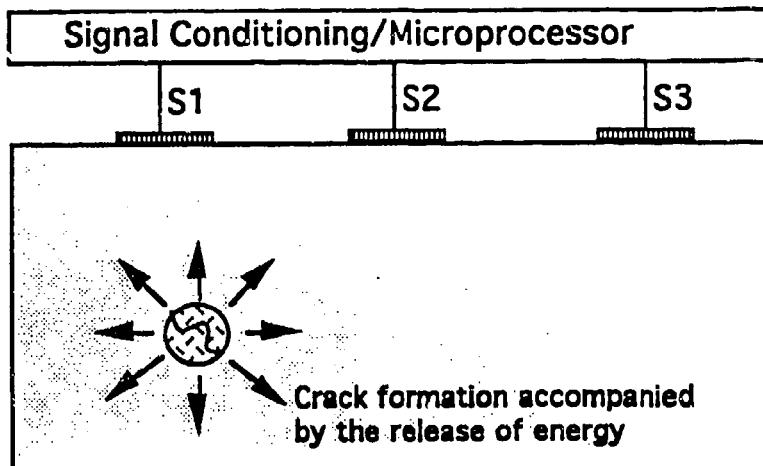
The technical approach to identification of aircraft failure is based on the use of microsensor technologies. Two approaches are considered: 1) real-time identification of the precursors to failure involving the detection of acoustic emission generated by material microcracking and 2) monitoring the time integrated stress of aircraft components and materials.

This section describes each of these technical approaches, the physical basis for acoustic emission detection in solids, and the use of microsensors derived from microelectromechanical systems (MEMS) technologies.

2.1 Component-Attached Microsensors for Acoustic Emission Detection

We envision two possible methods of implementing a direct sensing scheme for continuous in-flight monitoring of critical components. The first method, illustrated in Figure 1, would use multiple silicon die with dimensions of approximately 5 mm x 5 mm bonded directly onto components or embedded near the areas of probable crack formation or suspected high stress. The silicon die would contain the sensors which will sense the formation of microcracks and the necessary signal conditioning electronics. The sensors on the silicon die could include strain gauges, accelerometers, stress sensors, and acoustic emission sensors. While it is believed that acoustic emissions sensors would be most useful in detection of the high frequencies (> 100 kHz) associated with material cracking, other sensors could be used to provide additional information. Such multi-functional sensing capability would be useful in cross-correlation of sensory inputs and provide a built-in redundancy in a localized area for high reliability. Multiple silicon die could be inexpensively manufactured.

Microcrack initiation and growth generates a series of impulsive acoustic events that contain a very wide range of frequencies, often to several megahertz. The practical bandwidths achievable in most conventional sensors and macro-strain gauges is usually below 100 kHz and presents both a sensitivity and frequency limitation in this application. Micro-sized acoustic emission sensors can however be fabricated on the silicon die with sufficient bandwidth and high frequency response to detect the impulsive acoustic events generated by microcrack development.



**Figure 1. General Sensing Principle Used
To Detect Crack Formation In Solids**

Acoustic energy is released due to material cracking and microcrack growth. This sound is transmitted through the material medium and detected by acoustic emission sensors either embedded within the solid material (not shown) or mounted on the material surface (as shown). The use of multiple acoustic emission sensors allows triangulation methods to be used in determining the location of the crack. The amplitude of the acoustic emission signature determines the severity of the crack.

Stress sensors, strain gauges, and accelerometers would be most useful for long term routine monitoring of the stress-strain loading cycles to which the component is subjected. If this information can be kept up-to-date over the lifetime of the component, then the condition of the component could be inferred with a high degree of confidence. Acoustic emission sensors would be most useful for the real-time detection of crack formation.

The major technical considerations in this sensing approach are described below.

- The silicon based sensors must be rigidly attached to the critical components. Anodic low-temperature bonding techniques of the chip package (or silicon material) to steel components need to be considered.
- Piezoelectric thin films deposited on silicon wafers and microstructures are suggested for multi-functional sensing. Microsensors to be designed include 1) accelerometers, 2) strain gauges, 3) stress gauges, and 4) acoustic emission sensors. Both real-time and life-cycle sensors can be implemented.
- On-chip signal processing electronics must be integrated with piezoelectric microsensors. Operational amplifiers to be used must transduce charge into an output voltage with minimal signal loss due to parasitic capacitive electrical interconnection.

- Custom on-chip integrated circuits must be designed as needed for the special signal processing requirements associated with piezoelectric thin film sensors and the specific high noise ambient of the aircraft. The use of reference cancellation circuitry is needed to reject unwanted vibration, engine noise, and ambient temperature fluctuations.
- The piezoelectric microsensors must be tested in simulated crack generated environments. Signal-to-noise characteristics for sensors need be established along with interpretation of signal waveforms.

2.2 Physical Basis for Crack Detection Using Acoustic Emission Sensors

The use of acoustic emission sensors in detecting crack formation and growth has routinely been used over the past twenty years in three application areas: structural testing and surveillance, process monitoring and control, and materials characterization. Acoustic emission monitoring methods are widely used in non-destructive testing (NDT) and non-destructive evaluation (NDE) applications. A discussion of acoustic emission wave detection in thin plates is given in Ref. 6. A general review of the instrument science and technology developments is given in Ref. 7.

Of the large number of sensing approaches which might be considered for the monitoring acoustic emission, piezoelectric materials are among the most widely used primarily for their excellent signal-to-noise characteristics [8]. The physical basis for the detection of acoustic energy release in solid materials by surface-mounted piezoelectric transducers has been previously developed by Rose [9]. The theory for piezoelectric sensing devices for acoustic emission monitoring will not be repeated in this work.

2.3 Microelectromechanical Systems (MEMS) Technology

The approach taken in this paper is uses intelligent sensor chip manufactured by microelectromechanical systems (MEMS) techniques. MEMS technology has its origins in the late 1980s and has demonstrated commercial devices only within the last two years. These manufacturing methods are based on extensions of the integrated circuit process and are primarily used to make sensors, actuators, and microstructures compatibly with on-chip electronics. The premise of MEMS technology is that miniature intelligent systems can be produced in large bulk quantities through methods similar to those used in the manufacture of commodity memory chips.

The manufacturing methods include those techniques used in the production of either silicon or gallium arsenide integrated circuits with additional process enhancements

which include solid-state micromachining and specialized packaging methods. Through selective ordering of process modules using thin film deposition, lithography, etching, and wafer bonding techniques, several MEMS devices have been demonstrated. One promising commercial device currently marketed by Analog Device, Inc. (Norwood, MA) is a smart automobile airbag deployment accelerometer with on-chip signal conditioning electronics.

Figure 2 shows an example of MEMS technology applied in the formation of a high frequency acoustic emission sensor. The pressure sensor chip shown makes use of a 1.0 μm -thick, deformable, silicon nitride membrane with a deposited 3600-Å piezoelectric thin film. As the membrane deflected due to incident acoustic energy, a charge is produced by means of the piezoelectric effect. This charge is directly coupled on an on-chip MOS transistor and converted into an analog voltage. The amplitude of the voltage waveform is directly proportional to the acoustic vibration [10].

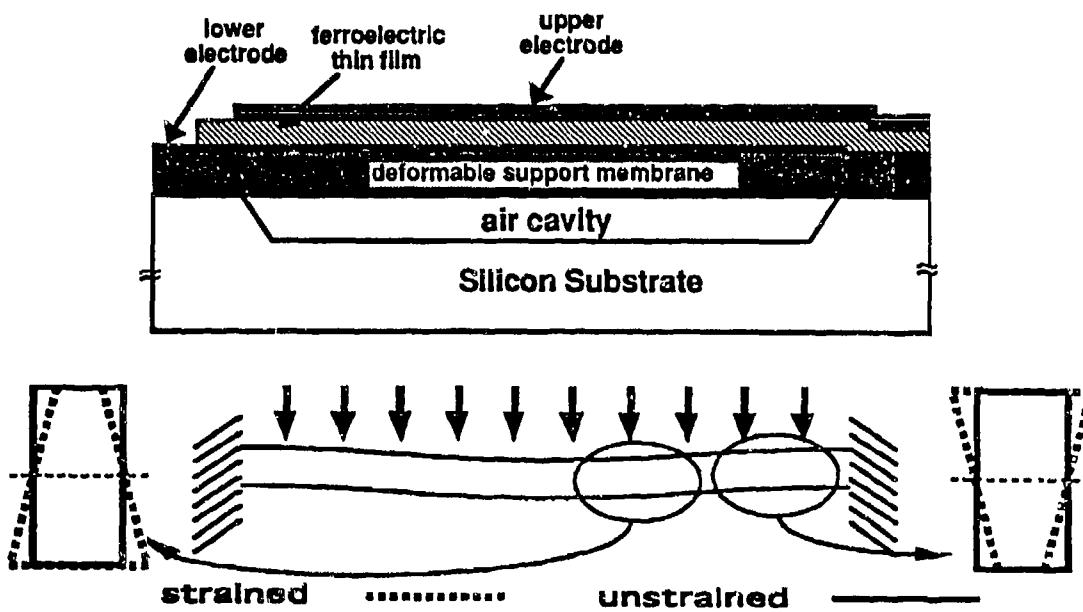


Figure 2. Cross-Section of a MEMS Acoustic Emission Sensor

A 1.0 μm -thick silicon nitride membrane supported at its edges deforms in response to the force accompanied by the release of acoustic energy. The movement of the silicon nitride membrane stresses an overlaying piezoelectric thin film capacitor which produces a charge in proportion to the mechanical excitation. This charge is transduced to an on-chip amplifier and converted into a usable voltage waveform. Typical membrane diameters are between 20 to 500 μm and can be adjusted to meet a particular frequency response of interest.

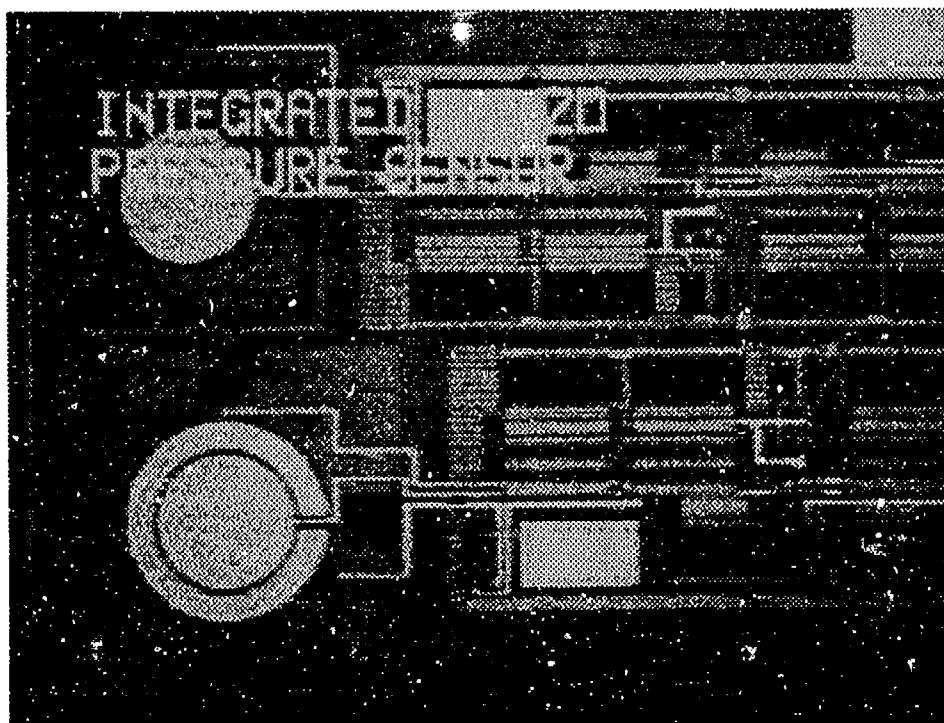


Figure 3. Optical Photograph of an Integrated Acoustic Emission Sensor

This device consists of a physical piezoelectric sensor directly connected to an on-chip operational amplifier. Device was fabricated at the University of Minnesota Center for Microelectromechanical Systems [10].

2.4 Comparison with Other Approaches

Several studies have previously been carried out in the use of acoustic emission sensors in detecting crack formation in aircraft materials. Horoschenkoff and Wittman used large piezoelectric sensors to detect delamination initiation in carbon fiber composites [11]. Bardenheier has studied acoustic emission in polymer composite materials [12]. Scala, et al., have suggested acoustic emission sensors for detecting fatigue crack propagation as a means of assessing structural integrity of aircraft [13, 14].

These studies have demonstrated encouraging results in acoustic emission monitoring of microcrack growth. Most of the above studies have used large, bulk-grown, piezoelectric acoustic emission sensors which generally have a limitation in frequency response and signal loss due to parasitic interconnection capacitance between the sensor and associated electronics. In the above studies, active noise cancellation methods have not been performed and consequently one of the difficulties previously encountered has been the inability of these approaches to reject false signals such as those due to engine noise and

surfaces moving over one another. Furthermore, these sensors are not responsive to frequencies above 100 kHz.

The major difference between the approach described in this study and those previously carried out is the use of microsensor technologies. The most significant advantage of microsensors being applied for crack detection and fatigue monitoring lies in 1) their extremely small size, 2) high-frequency response, and 3) low system weight.

3. CONCEPT FEASIBILITY

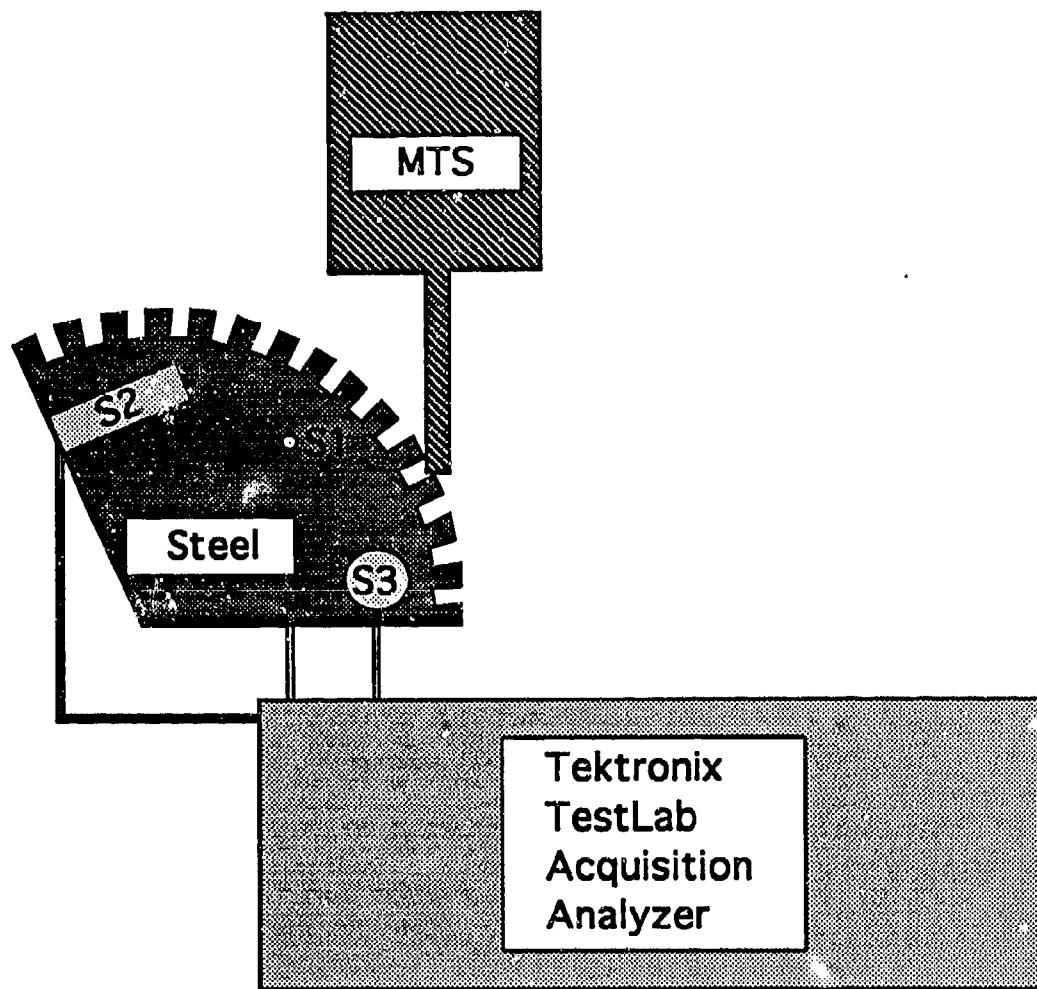
Two concept demonstration experiments were carried out to evaluate the feasibility of the sensor concepts proposed. The first experiment assesses the performance of a simple acoustic emission sensor in detecting material cracking. The second experiment assesses the use of an integrating fatigue sensor which counts the number of material flexures with a certain threshold amplitude.

3.1 Feasibility of Acoustic Emission Sensors

A simple experiment was carried out to determine the feasibility of microcrack detection using acoustic sensors. The test apparatus used is shown in Figure 4. Three sound sensitive devices were mounted at different locations on a 3/8-inch stainless steel plate with machined gear teeth. A MTS Compressional Force Tester was used to fracture several of the teeth through the application of a commanded force. Several electronic data acquisition units were used to monitor and record the analog waveforms generated by the piezoelectric sensors.

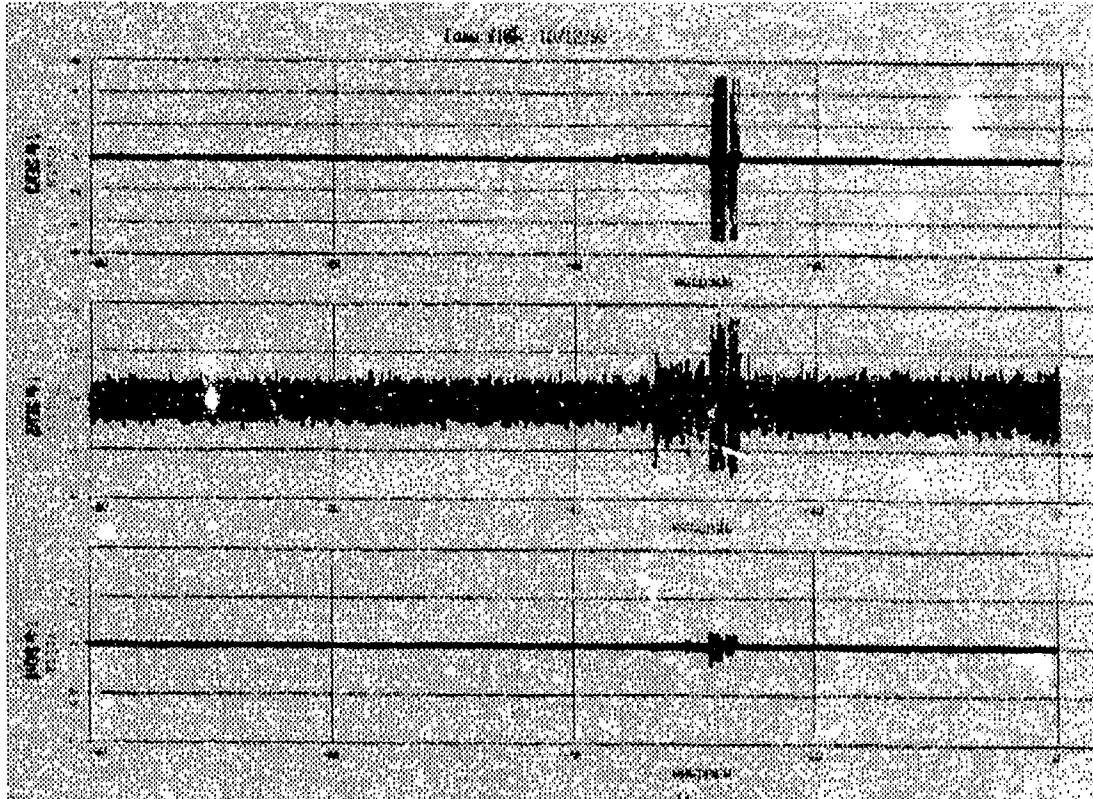
Sensor S1 is a MEMS device with a diaphragm diameter of 400 μm . Sensor S2 is a Kynar thin film and sensor S3 is a commercially available APC 12-400-850 pressure sensitive piezoelectric disk. Each of the three sensors was connected to an external amplifier and with an output voltage gain arbitrarily adjusted to be in the range of 0 to 5 V.

Figure 5 shows three waveform traces generated by the acoustic sensors. All three sensors recorded the acoustic emission characteristic of the metal tooth fracture. While varying in signal-to-noise, the three output waveforms demonstrate a basic acoustic emission detection capability.



**Figure 4. Test Set-up Used To Evaluate the Performance
of Three Acoustic Emission Sensors**

A MTS Compressional load tester is used to fracture a tooth from a stainless steel gear. Three acoustic emission sensitive devices are located on the steel to detect the acoustic emission.



**Figure 5. Amplified Analog Voltage Waveforms of the
Three Sensors Used To Monitor the Acoustic Emission
Released Upon Fracture of the Gear Tooth**

One of the principal advantages of microfabricated pressure sensors lies in the ability to make these devices extremely small. Small size implies a high mechanical resonant frequency and consequently, high frequency response. This is one of the primary limitations of sensors fabricated by large non-integrated circuit processing methods. In order to successfully detect the extremely high frequencies generated in microcrack formation, the fabrication of variable area microsensors is proposed. This scheme allows frequency response tailoring through simple variation in size. Based on the responses of individual size elements in an array, it should be possible, in principle, to measure the acoustical frequency spectrum associated with crack formation and possibly reject unwanted slowly varying signals such as those associated with temperature change. Although decreasing the area of a piezoelectric signal diminishes the signal produced per unit energy, the multiple connection of individual sensor capacitors in parallel can be used

to enhance the overall charge produced by the piezoelectric effect. Figure 6 shows a plot of mechanical resonant frequency versus diaphragm diameter [10].

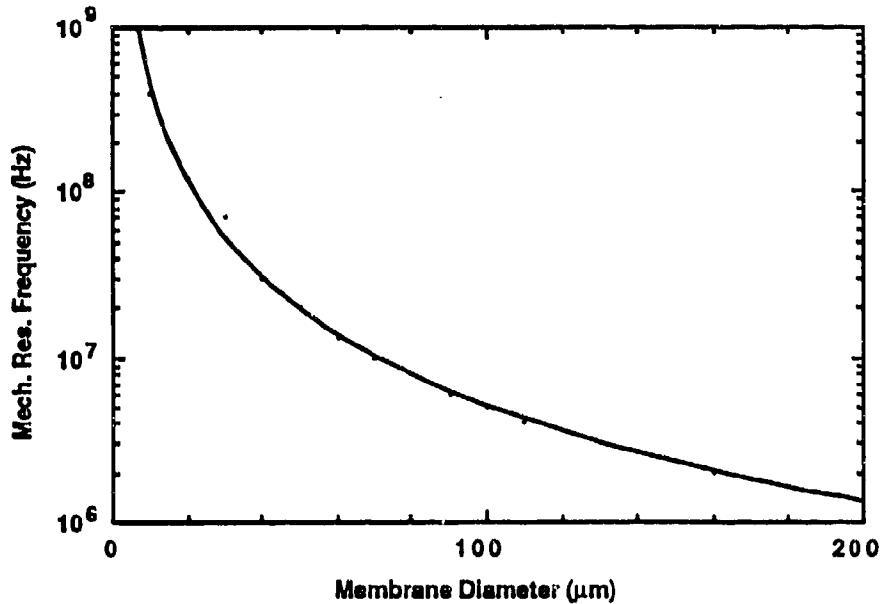


Figure 6. Calculation of Mechanical Resonant Frequency Versus Membrane Diameter

This graph points out the versatility of tailoring the frequency response through adjustment of diaphragm size.

3.2 Feasibility of Fatigue Sensors

A preliminary experiment to determine the feasibility of an integrating fatigue sensor was carried out using a thinned silicon wafer substrate (100 μm -thick) with an on-chip deposited PZT piezoelectric strain gauge. The strain gauge dimensions were approximately 1 cm x 4 cm in this diagnostic device. The base of the silicon wafer was mounted with epoxy on a flexible graphite fiber sheet supplied by McDonnell Douglas.

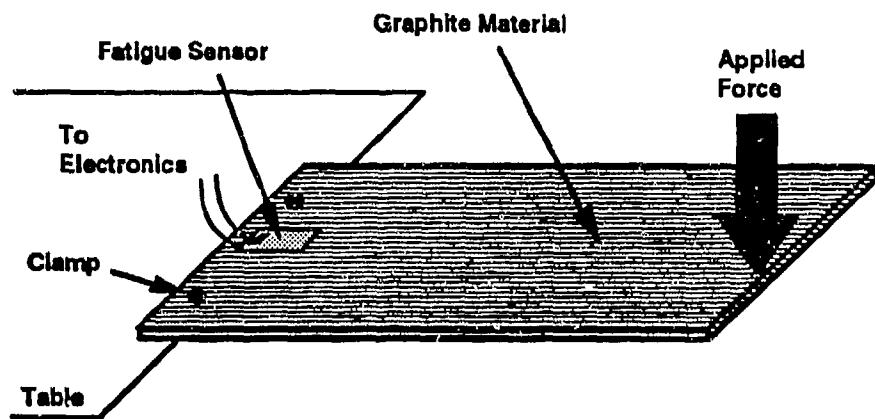


Figure 7. Test Set-up for Fatigue Monitoring

A piezoelectric microsensor is mounted on a graphite cantilever. The flexure of the cantilever induces a piezoelectric voltage. Threshold detection techniques are used to record the number of deflections which exceed a certain preset value.

A simple test was used to explore one possible mode of operation for this device. As shown in Figure 8, the output of the piezoelectric sensor was connected to the input of a low-noise differential amplifier. A simple threshold detector/trigger circuit was constructed to record voltage excursions above 0.2 V. The number of these excursions was counted by a simple digital memory cell.

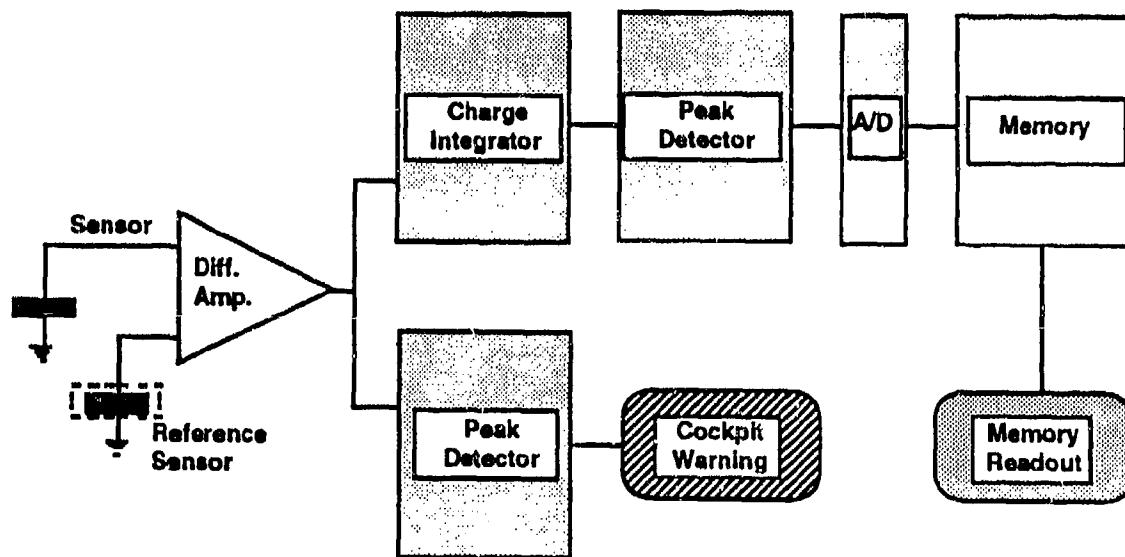


Figure 8. Signal Conditioning Circuitry for the Integrating Fatigue Sensor

4. KEY TECHNICAL CHALLENGES

There are several technical challenges needed to be addressed if either of the preceding sensor approaches is to be undertaken. This section outlines some important and major considerations.

4.1 Operational Mode of MEMS Devices

Several operational modes might be considered for the sensors proposed in this work. Both real-time and ground-based acoustic emission sensors and fatigue monitoring devices might be considered. Real-time sensors have the disadvantage in that both highly accurate and reliable sensor determinations must be made (such as for the pilot to bring the aircraft down). Therefore warning devices with some form of fail-safe or redundant features must be incorporated into these designs. Considerable progress has been made in this respect as demonstrated in self-calibrated automobile air bag deployment devices, for instance.

Perhaps a more important mode of operation is in ground-based maintenance. Here sensors are interrogated on the ground. Electronic waveforms characteristic of the integrated acoustic emission and/or fatigue seen by individual components are read out by ground service personnel. This mode of operation has the desirable feature of identifying which parts need to be replaced but has the disadvantage of not being able to detect precursors to catastrophic failure in-flight.

4.2 Sensor Packaging and Attachment

Each of the above two sensing approaches still requires a package and appropriate mounting to the surfaces being sensed. The packaging problem is a general one for all types of microsensors: only one desired variable is to be detected by the active element with all other variables rejected. For the acoustic emission sensor, the package must not significantly attenuate the sound. For the fatigue sensor, appropriate bonding to material surfaces is necessary. Package to component bonds must not delaminate during flight operation.

The mounting of silicon chips containing electronics and perhaps sensors is not expected to be an insurmountable technical problem. Silicon die are routinely mounted in hermetically sealed metal packages which are in turn mounted in a wide variety of ways and subjected to many different environments including some with high operating temperatures, high stresses, and large accelerations. In such environments, the circuits on the silicon chip

have long operating lifetimes. A large body of technical knowledge regarding the mounting of silicon die exists.

4.3 Noise Cancellation and Cross Sensitivity

An important question is whether or not the acoustic emissions from crack formation can be distinguished from ultrasonic signals generated by normal operation of the aircraft systems during flight. In particular stick slip friction between moving parts and various wear mechanisms are expected to generate acoustic noise similar in bandwidth to the acoustic emissions from crack formation. All aircraft have numerous acoustic, vibration, and electromagnetic interference noise sources. These sources of noise potentially make the unambiguous detection of crack growth and fatigue a difficult challenge. Therefore in this work the use of a non-attached reference sensor is proposed to detect these unwanted noise sources. To overcome the problem, both the actual sensor and reference sensor are to be connected to the inputs via a differential pair amplifier with high common-mode rejection ratio. In the past this has been a difficult task due to mismatches in the large sensor and amplifier input transistors and their electrical interconnection. Because both the sensor and amplifier are located on the same MEMS chip, parasitic effects such as those due to stray interconnection capacitance and electromagnetic interference pick-up are minimized. The use of differential amplifier cancellation techniques is also generally effective in the minimization of slowly varying changes in temperature and ambient pressure.

Research will be required to determine if the signals from crack formation have any unique features in the time domain, frequency domain, or spatial domain (location of probable crack site compared to location of noise emanations) that can be exploited to distinguish it from all other acoustic signals in the environment.

A related issue in the question of optimum placement of the sensor on the component to be monitored. It is expected that some sensor locations on a particular component will provide greater sensitivity to crack-related acoustic signals than other locations. In addition there may be locations that minimize the sensitivity of the sensor to signals generated by stick-slip friction or wear mechanisms. Modeling studies using finite element methods coupled with experimental tests are expected to be very useful in addressing these questions.

4.4 Positioning, Signal Interpretation, and Reliability

One of the key issues in these sensing approaches concerns reliability. In particular, the sensors used must operate reliably for periods in excess of ten years without unwanted change in their performance. This is necessary due in part to the required placement of the sensors in inaccessible regions or incorporation of the sensors into the manufactured components (such as embedding). This requirement obviously places significant demands on the design of the sensor. Furthermore, the sensors must have either a higher fatigue life than the parts being monitored or be placed in regions where smaller deflections are recorded as being proportional to the larger deflections in the regions of highest stress. Placement of these sensors and the interpretation of their signals over time is therefore an important consideration and one which requires substantial study.

4.5 Power Delivery and Signal Readout

The MEMS devices suggested in this work require external power and electrical readout of information. This is a difficult task in helicopter rotor crack detection. While there is a 24-V dc power line available on the helicopter shaft, the use of slip-rings, energy storage devices, and some form of superimposed voltage waveforms may be necessary for both sensor power delivery and signal readout. The combined use of an electrical conductor line with superimposed dc (source) and ac (signal) components is suggested for this application. Several non-contact methods of signal transfer such as via radio, light (such as LEDs), or acoustic waves are possible, but each method has its own difficulties. Research will be required to delineate the magnitudes of the problems and to assess the possible solutions. This obviously requires more consideration than possible in this present work.

In monitoring fatigue in multiple sensor locations (> 1000), significant total system weight would be added to an aircraft if each sensor contained one or more wires. Therefore some form of local multiplexed networks is suggested if several hundreds of sensors are to be used such as along the fuselage. These locally multiplexed networks would be connected to a central microprocessor which makes interpretations based on the multiple sensory inputs processed over time.

Figure 9 shows a suggested approach in the fabrication of a smart MEMS module (contains multi-functional sensors and integrated circuits on a common substrate). The key feature of this approach is to maximize system functionality and minimize the number of interconnection points. Such a multi-functional device has built-in temperature,

acceleration, and ambient pressure referencing capability. Other silicon microsensors might also be added as needed.

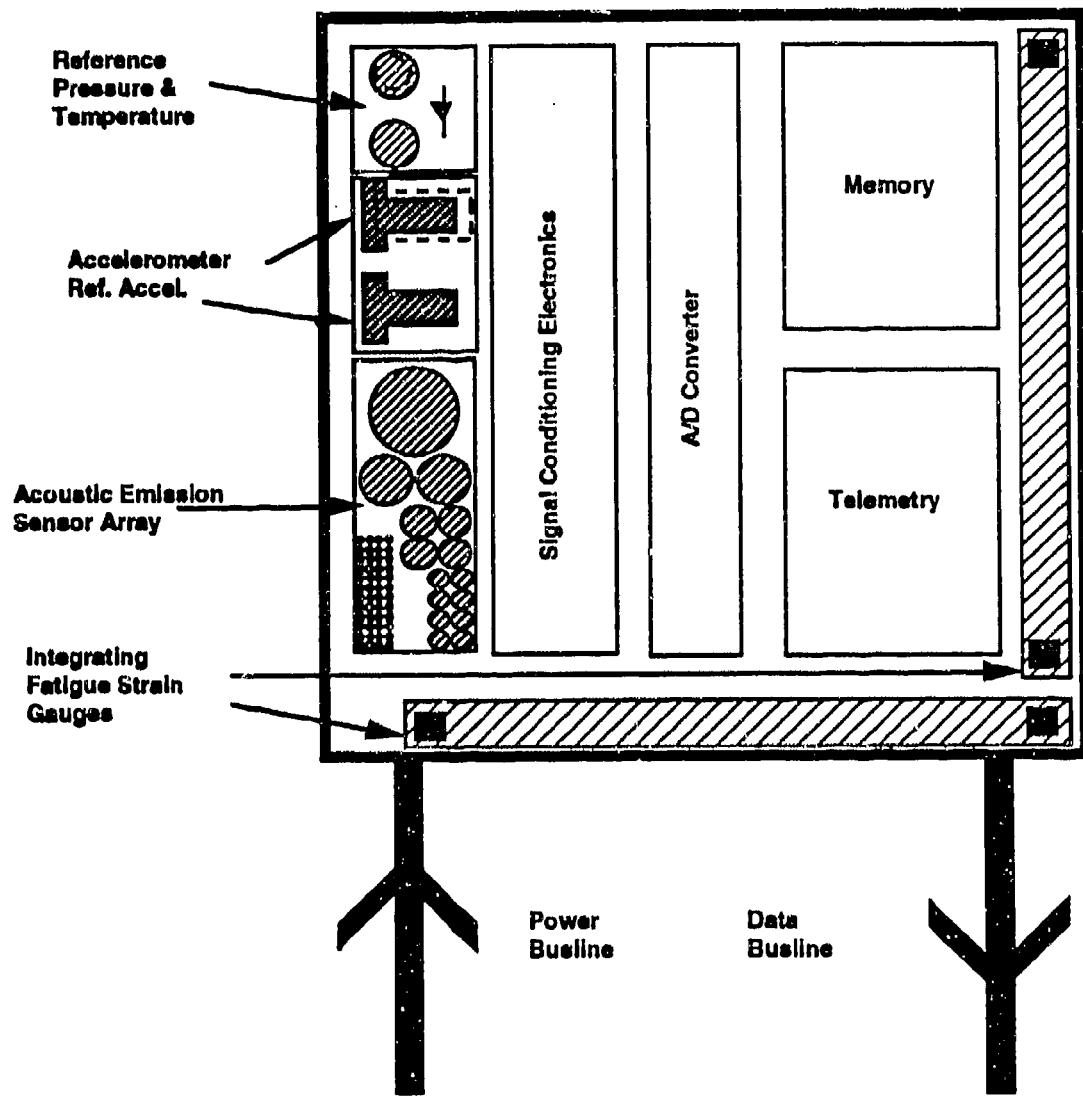


Figure 9. Multi-Sensor Concept for Recording Both Acoustic Emission and Fatigue

Several other sensor inputs are used to monitor ambient pressure and temperature changes. Cantilever beam accelerometers are used to cancel component vibration.

5. CONCLUSIONS

The use of microelectromechanical systems for monitoring acoustic emission associated with crack formation in aircraft components represents a new approach to a problem recognized for over the past 30 years. The main aim of this work is to improve reliability of aircraft components and possibly extend their usable lifetime with only the necessary replacement of parts taking place over the service lifetime of the aircraft. There are obviously many reliability and test issues which still need to be addressed.

The main conclusions of this work are summarized below.

- The application of newly developed microelectromechanical systems (MEMS) technology to the monitoring of acoustic emission associated with crack formation in aircraft components.
- The application of integrating fatigue microsensors based on MEMS technology to determine fatigue in aircraft components.
- The application of piezoelectric microsensors to acoustic emission monitoring where the small size afforded by MEMS technology leads to an enhancement in the high frequency response of these sensors.
- The application of piezoelectric strain gauges deposited on thinned silicon wafers to measure component flexure in regions near points of high stress.
- The use of intelligent multi-functional chips containing a variety of physical sensors to monitor not only acoustic emission and fatigue but also general variables of interest such as absolute pressure, acceleration, and temperature.
- The use of multiple, inexpensive, small-size sensory modules to collect materials flight performance data. This information can be used in real-time or interrogated after flight by ground service personnel.

ACKNOWLEDGMENTS

The author would like to thank Dr. Robert Pohanka of the Office of Naval Research for initially bringing the aircraft reliability problem to his attention. Helpful conversations with IDA staff and mentors, DSSG colleagues, and University of Minnesota students are gratefully acknowledged. The author would like to thank Professors W. Robbins, A. Erdman, and J. Vogel of the University of Minnesota for their interest in the crack formation problem and assistance in the initial feasibility experiment. Concept demonstration devices were fabricated in the University of Minnesota Center for Microelectromechanical Systems.

REFERENCES

1. See for example the yearly *IEEE Proceedings on Microelectromechanical Systems Workshop*, 1986-1993.
2. See for example *Microsensors*, (R.S. Muller, R.T. Howe, S.D. Senturia, R.L. Smith, and R.M. White, eds.) IEEE Press, New York 1990.
3. K.D. Wise, "Integrated Silicon Sensors: Interfacing Electronics to a Non-Electronic World," *Sensors and Actuators* 3, 229 (1982).
4. S. Senturia, "Microsensors vs. Integrated Circuits: A Study in Contrasts," *IEEE Int. Electron Devices Mtg.*, Washington, D.C., 1989.
5. R.T. Howe, "Surface Micromachining," *J. Vac. Sci. Technol.* B6, 1809 (1988).
6. D.M. Egle and A.E. Brown, "Considerations for the Detection of Acoustic Emission Waves in Thin Plates," *J. Acoust. Soc. Am.* 57, 591 (1975).
7. C.B. Scruby, "An Introduction to Acoustic Emission," *J. Phys. E: Sci Instrum.* 20, 946 (1987).
8. J. Bennett, J. Paul, R. Jones, and A. Goldman, "A preliminary Study on Damage Detection Using Piezoelectric Film," technical report, Australian Department of Defense, AR-006-620, 1991.
9. L.R.F. Rose, "The Stress-Wave Radiation from Growing Cracks," *Int. J. of Fracture* 17, 45 (1981).
10. P. Schiller C. Ye, T. Tamagawa, and D.L. Polla, "Design and Process Considerations for Ferroelectric Film-Based Piezoelectric Pressure Sensors," *4th International Symposium on Integrated Ferroelectrics*, Monterey, CA, Mar. 1992.
11. A. Horoschenkoff and E. Wittmann, "The Use of Piezoelectric Sensors to Detect Delamination Initiation in Carbon Fibre Reinforced Composites," technical report, Deutsche Aerospace, N-9232740, 1991.
12. R. Bardenheir, "Acoustic Emission Monitoring of Polymer Composite Materials," NASA Technical Memorandum, NASA TM-76523, 1981.
13. C.M. Scala, S.J. Bowles, and I.G. Scott, "The Development of Acoustic Emission for Structural Integrity Monitoring of Aircraft," Australian Department of Defense, AR-004-585, 1988.
14. C.M. Scala and S. McK. Cousland, "Acoustic Emission during Fatigue Crack Propagation in the Aluminum Alloys 2024 and 2124," *Materials Science and Engineering* 61, 211 (1983).

**B. AN ASSESSMENT OF THE STATE-OF-THE-ART IN
CONSTITUTIVE MODELING AND HYDROCODES FOR
BALLISTIC IMPACT AND BLAST WAVES**

**David L. McDowell
Georgia Institute of Technology
Atlanta, Georgia**

**Peter W. Voorhees
Northwestern University
Evanston, Illinois**

EXECUTIVE SUMMARY

In 1980, a National Materials Advisory Board panel reviewed the current state-of-the-art of modeling and understanding of the behavior of materials under the high strain rate loading typical of projectile penetration and blast wave impact. This panel endorsed a concurrent modeling and experimental approach which has strongly influenced the development of the field since their report. Given the rapid progress made since this study in our understanding of damage mechanisms in materials at high strain rate and the exponential growth of computing power and numerical simulation schemes, we have assessed the current state-of-the-art and have made recommendations for further study.

The importance of modeling and simulation in the development of new armor and anti-armor has never been more acute. The financial limitations accompanying the end of the cold war have drastically impaired the feasibility of full-scale testing of new armor and anti-armor systems. This is particularly troubling as it is well-known that laboratory scale testing can lead to qualitatively incorrect conclusions when compared to full-scale testing of a similar system. Thus, modeling and simulation appear to be the ideal vehicles by which this problem can be addressed at reasonable cost.

Hydrocodes have developed to a point where solving the field equations governing the temporal and spatial evolution of force, displacement and temperature at the high rates associated with impacts is relatively straightforward. The "weak link" of hydrocodes is the constitutive models used to relate these quantities as well as to determine the extent of damage or vulnerability. In many cases, unknown constants in these constitutive relations are chosen to ensure that the calculation agrees with a particular impact experiment. Thus, it is difficult to use most existing models to extrapolate beyond the materials and small-scale experiments used to calibrate a given hydrocode.

To enable modeling and simulation to fulfill their anticipated central role in the development of armor and anti-armor systems and concepts, it is necessary for micromechanically based constitutive laws for deformation and damage be employed in existing hydrocodes. A general thermomechanical framework is outlined within which these laws may be considered.

AN ASSESSMENT OF THE STATE-OF-THE-ART IN CONSTITUTIVE MODELING AND HYDROCODES FOR BALLISTIC IMPACT AND BLAST WAVES

INTRODUCTION

Many important applications involving blasts and projectile penetration of interest to the U.S. military have been buoyed technically by an aggressive ballistic testing program. In its 1980 assessment of the start-of-the-art in dynamic material behavior, characterization and modeling, a National Materials Advisory Board (NMAB) panel formed by the National Academy of Sciences recommended that simulation, ballistic testing and characterization of material response should continue to be performed concurrently with increasingly active communication between these efforts (NMAB, 1980). This recommendation grew out of the panel's assessment that a fundamental characterization of material failure mechanisms was significantly lacking, so too much reliance on simulation would not prove reliable. In addition to material deformation and failure studies, the development of Lagrangian and Eulerian hydrocodes for simulation of blast wave and penetration events has proceeded with vigor within the last 15 years. Deformation and failure mechanisms associated with impact depend significantly on impact velocity and properties of the penetrator and target materials within certain velocity regimes. The field of impact mechanics has followed certain guidelines established three decades ago regarding treatment of the hydrodynamic and plastic flow aspects of impact. Within the past 15 years, a number of different damage mechanisms have been identified such as nucleation and growth of voids, microcracks, adiabatic shear bands, solid state and solid-liquid phase transformations, etc.; moreover, these various mechanisms have been modeled with varying degrees of rigor and detail. At the same time, hydrocodes have been extended to include many of these mechanisms, as well as additional material erosion effects which extend range of applicability and accuracy of the codes for certain applications. In effect, some of these new developments, as well as procedures traditionally adopted in these codes, amount to the introduction of additional constitutive relations which may be related more to computational efficacy than realistic material flow and damage evolution.

The purpose of this study is to examine the progress made in identification of material deformation/failure mechanisms and models for these processes as implemented in current hydrocodes, concentrating particularly on the decade since the 1980 NMAB report cited earlier. Recent progress in our understanding of deformation and damage mechanisms under shock loading will be reviewed, as well as the state of model development in representing these mechanisms. Comments and recommendations will be made regarding the use of hydrocodes of impact simulation and vulnerability/lethality analyses. Some suggestions will be made regarding the use of internal state variables to represent deformation and damage processes, as well as a corresponding thermodynamically consistent framework for inclusion of these effects within the context of the impact problem. It will be argued that an internal state variable framework offers a common format to include micromechanical models for the various coupled deformation and failure modes.

1. OVERVIEW: APPLICATIONS OF INTEREST

In this report, we will consider dynamical response of structures subjected to blast wave or penetrator impacts, concentrating on metallic targets. To date, metallic targets form the basis of most reported developments in characterization and modeling regarding blast wave damage, kinetic energy (KE) penetrators and shaped-charges. More recently, microcracking and comminution models have been developed for confined ceramic armors (Shockey et al, 1990). Moreover, layered composite armors (cf. Radin & Goldsmith, 1988) and active and reactive armors (Zaloga, 1987) have become intense areas of research, with the aim of defeating KE penetrators and shaped-charges.

Before discussing specific applications, a brief review of the physics of stress wave effects arising from impacts and blasts is in order. For penetrator impact, one may characterize the problem in terms of the velocity of the penetrator, V , relative to the velocity of sound in the target material, c . A hypervelocity impact occurs when $V/c > 1$. This is the so-called ideal or hydrodynamic penetration regime for long-rod penetrators (e.g., shaped-charge jets, electro-magnetically (EM) gun fired KE projectiles, ballistic missile boosted projectiles). The velocity of sound for structural metals is in the vicinity of 4-5 km/sec. For hypervelocity impacts, the compressive shock wave produces extremely high peak pressures (100 to 200 GPa), but does not significantly outpace the front of the eroding penetrator. Hence, the material is disrupted and damaged during the penetration over a fairly local scale, typically on the order of only a few penetrator diameters. Only very high frequency, low amplitude modes of structural vibration are excited. As a result, the structural modal frequency response and structural constraints are of minimal importance. Reflected waves result in tensile stresses which induce additional damage in the form of material microcracking, spallation, etc. and act to reduce the pressures rapidly after penetration.

Due to the generation of such high pressures under hypervelocity impact conditions, the role of deviatoric (shear) stresses in material deformation and damage behavior is typically very small compared to that of the hydrostatic stress. Hence, an equation of state relating the pressure to the specific volume of the material is extended from the case of fluids and experimentally characterized under shock conditions for metals. Under these conditions, the impact may be approximately modeled as interpenetration of two fluids, with the density of projectile and target materials playing a key role in defining depth of penetration, penetration velocity, etc. Material strength and failure effects are held

to be of secondary importance, with the exception of finite width and thickness effects which promote progressive damage due to stress wave reflections.

Hypervelocity impacts result in very significant adiabatic temperature rises, with average temperatures ranging from 0.2 to 0.5 of the melting point. Peak local temperatures may induce melting. Strains are typically very significant, on the order of unity, with peak local strain rates on the order of 10^6 s^{-1} to 10^7 s^{-1} . Typically, very high strains are locally concentrated in the vicinity of the penetrator front and flank, as material is deformed, fractured, eroded, etc. ahead of the moving penetrator. Time scales for penetration and associated material deformation and failure mechanisms are on the order of microseconds. Very high velocity hypervelocity impacts ($> 12 \text{ km/s}$) can result in extensive melting and vaporization of material. Such velocities are not achieved even by present gun-fired or shaped-charge penetrators.

In contrast, hypovelocity ($V/c < 1$) impacts are characterized by the propagation of a shock wave across the target material thickness and reflection well before penetration, depending on the impact velocity. In this case, tensile stresses induced by wave reflections may assist the penetration/cratering process and often result in a more diffuse distribution of deformation and damage. For hypovelocity impacts, the strength and failure properties of both the penetrator and target materials influence penetration significantly, with an increasing influence at lower velocities. As impact velocity decreases, structural constraints, flexibility and modal response become increasingly influential and interact with the effects of the stress waves generated by the impact.

Typical ordnance velocity ranges are from 1.2 to about 1.8 km/sec. Within this range of impact velocity, there are substantial effects of material resistance associated with elastic-plastic flow and material damage, abrasion, and erosion processes. At higher velocities, from approximately 2 km/s to 4 km/s, there is a transition regime where the penetration problem is decreasingly influenced by material strength and failure behavior. For ordnance velocities, peak pressures are much lower, on the order of 20 to 40 GPa, and rapidly subside to the order of the material yield strength after multiple stress wave reflections. Temperatures are typically on the order of 0.2 to 0.3 of the melting point, with average strains and strain rates on the order of 0.3 and 10^4 s^{-1} to 10^5 s^{-1} , respectively.

For all velocity regimes, it is important to note that strains on the order of unity and strain rates on the order of 10^6 s^{-1} to 10^7 s^{-1} are typically generated in the target material near the tip of the projectile. Moreover, peak temperatures can be a significant fracture of the melting point and are very difficult to measure; often they are inferred from

microstructure morphology changes after the event. Table 1 presents typical target pressure, temperature, strain and strain rate responses for a range of impact velocities.

Table 1.
(NMAB Report, 1980)

	Pressure GPa	Homologous* Temperature	Strain	Strain Rate s ⁻¹
Shaped-Charge Jet (3–10 km/s)	Peak ~ 200 Avg. ~ 20	Peak > 1 Avg. ~ 0.5–0.7	>10	Peak ~ 10 ⁶ –10 ⁷ Avg. ~ 10 ⁴ –10 ⁵
Self-Forged Fragment (1.5–3 km/s)	Peak ~ 40 Avg. ~ 20	Peak ~ 0.5–0.8 Avg. ~ 0.2	Peak ~ 2 Avg. ~ 0.7	Peak ~ 10 ⁶ –10 ⁷ Avg. ~ 10 ⁴
Fragmentation (1.3–3 km/s)	Peak ~ 30 Avg. ~ 2	Ductile ~ 0.3–0.5 Brittle ~ 0.1	Ductile ~ 0.5–1.5 Brittle ~ 0.1–0.2	Peak ~ 10 ⁶ –10 ⁷ Avg. ~ 10 ⁴

Damage due to blast waves produced by detonation of high explosive (HE) near structural surfaces such as shells is highly dependent upon structural constraints since the stress waves have ample time to traverse the shell thickness compared to the period of structural oscillations. In fact, it is commonly observed that damage progresses in "surges" corresponding to reflected stress waves focusing on preferred sites during large deformations (cf. Bamann et al., 1993). Blast wave damage of shells is often dominated by the geometric nonlinearity of shell deformation; strains are relatively small, nominally on the order of 0.1, compared to penetrator impacts. Pressures generated by the shock waves are typically on the order of the material yield strength; accordingly, material deformation and failure modeling are of prime importance. Relatively low frequency vibratory response modes are elicited in the structure, with time scales on the order of milliseconds for the cumulative damage process.

The blast waves due to HE detonation at some stand-off distance are transmitted through a fluid to the target. The shape of the charge and orientation with respect to the surface varies the effectiveness of the blast through focusing effects; for example, cylindrical charges, end-on, are most effective on the target when detonated within one blast radius (Goodman, 1982). A shock wave generated by collapse of the cavity created by detonation typically follows the initial shock wave, producing additional damage.

In all these cases, the profile of the shock wave is significantly affected by plastic deformation behavior, evolving material damage, viscous material response, etc., as evidenced by the effect of inelastic behavior on the shock Hugoniot curve. Accordingly, the shock impulse delivered to the material depends on these factors and the actual case is

far from that of an elastic stress wave problem. This necessitates numerical solution for essentially all problems of practical interest and, in fact, complicates interpretation of dynamic experiments which are used to determine material parameters. This is discussed in more detail in Sections 2 and 3.

1.1 Long Rod Kinetic Energy (KE) Penetrators

Cylindrical penetrators with a high length to diameter ratio (10-25) are quite effective in armor penetration. Such penetrators progressively erode at the tip during the penetration, leading to backward extrusion of penetrator and target material, while the rear surface moves with almost constant velocity. Extremely high pressures, inelastic flow and damage states are generated in the vicinity of the penetrator tip. The target material also undergoes extensive deformation, microdamage and erosion. Target hole diameters on the order of several projectile diameters (e.g., 5-10) are common. Typically, rolled homogeneous armor (RHA) is used, which is a military specification representative of 4340 steel alloys. A wide variety of penetrator materials have been used, including steel, tungsten, tantalum and depleted uranium (DU).

For subsonic velocities (Ferrari, 1988), it is desirable to have the maximum possible velocity with the smallest cross sectional area of impact to penetrate effectively, since the energy required for penetration is roughly proportional to the diameter of the penetrator. However, the energy made available by the gun to the projectile increases with the cube of the gun bore. Hence, the use of projectiles of caliber smaller than the gun barrel is quite effective; break-away sabots are used to encapsulate the penetrator and impart spin during the transit in the barrel. Figure 1 presents a typical geometry of a long rod penetrator with a sabot assembly.



Figure 1. Long Rod Penetrator, Aerodynamically Stabilized, With Sabot
(Weihrauch, 1987)

The physics of penetration depend intimately on impact velocity. For hypersonic impacts, in general, the deformation and damage mechanisms in the penetrator and target materials play a very strong role. Early hydrodynamic theories (cf. Tate, 1967) based on impact of interpenetrating fluids do not correlate the data. If the penetrator does not perforate the target, cratering or additional penetration will occur near the end of penetration due to interaction of reflected stress waves and transfer of momentum. Cratering phenomena are highly dependent on material failure properties.

For hypervelocity impacts, modified hydrodynamic theories have been introduced long ago which assume both the jet and the target materials are incompressible, perfect fluids. In these theories, Bernoulli's equation is augmented by pressure terms which account for both penetrator and target resistance to plastic deformation for the case of ductile metal targets, i.e.,

$$\frac{1}{2} \rho_t u^2 + R_t = \frac{1}{2} \rho_p (V - u)^2 + Y_p \quad (1)$$

where R_t and Y_p represent resistance pressures to plastic flow in the target and penetrator materials, respectively (cf. Tate, 1967, 1969). Naturally, the values of these pressures (particularly R_t) depend on material, geometry and projectile velocity since they reflect constraint on plastic flow in analogy to the Prandtl punch problem. Densities of the penetrator and target materials are subscripted by t and p, respectively. Equations are presented for the rate of target penetration and penetrator erosion during impact, i.e.,

$$\frac{u}{V-u} = \frac{1}{1-\mu^2} \frac{V - \mu \sqrt{V^2 + A}}{V - u} \quad (2)$$

where u is the velocity of the jet/target interface, and V is the penetrator velocity. It is argued that very little of the impact energy is absorbed by plastic deformation of the target material in the local vicinity of the penetrator, since the hydrostatic pressures are very large compared to the shear stresses (20 to 60 times the ultimate strength), but rather by extensive plastic flow in cratering (spallation, microfracture processes, etc.) and in plastic deformation away from the impact in the target after conversion of hydrostatic energy into kinetic energy of the target material.

Long rod penetrators at high velocity are modeled very similarly to shaped-charge jets using hydrodynamic theory, provided the impact velocities are sufficiently high

compared to sonic velocity. However, kinetic energy projectiles normally have much lower velocity, so that hydrodynamic analogies are not justified. Phenomenological equations have been developed based on extensions of the hydrodynamic theories for hypervelocity impacts to the subsonic case. For penetration of a continuous rod at high velocities (> 2 km/sec),

$$P = L \left(\frac{\rho_p}{\rho_t} \right)^{1/2} \left[1 - \frac{D}{L} \right] + \frac{1}{2} D \left[\left(\frac{\rho_p}{\rho_t} \right) \rho_p \frac{V^2}{2S} \right]^{1/3} \quad (3)$$

where V is the penetrator velocity, L is the penetrator length, D is the penetrator diameter, P is the penetration depth, and S is the "cratering resistance stress" (Charter & Orphal, 1991). This phenomenological equation interpolates between the hydrodynamic solution for large L/D , in which the classical first term is operative (velocity and target strength independent) to the $L/D = 1$ limit of an impacting sphere (second term) which is velocity dependent and engenders a "cratering" mode of damage at the penetrator front. This expression was shown to fit data rather well. Essentially, these two terms also represent phases of penetration, even for a long penetrator. Segmented long rod penetrators conceptually enhance penetration by producing N cratering events at the end of the penetration of each of N segments. For lower velocities (< 2 km/sec), the long rod solution is augmented by material resistance effects i.e.,

$$P = L \left(\frac{\rho_p}{\rho_t} \right)^{1/2} \left[1 - \frac{D}{L} \right] \left[1 - e^{-cv^2} \right] + \frac{1}{2} D \left[\left(\frac{\rho_p}{\rho_t} \right) \rho_p \frac{V^2}{2S} \right]^{1/3} \quad (4)$$

where C is another constant selected to represent material resistance.

As discussed later in this report, such approximate hydrodynamic relations must be regarded as qualitative guidelines compared to quantitative tools such as state-of-the-art numerical simulation packages.

Material failure mechanisms vary with target and penetrator materials and depend on impact geometry, projectile shape, etc. For conventional KE penetrators, mechanisms include brittle fracture/microcracking, spalling, punching (shearing, plugging), radial fracture, ductile penetration, petaling, and scabbing. These mechanisms are shown in Figure 2. The temporal occurrence of these failure modes depends largely on impact velocity. It is

widely recognized that material deformation and damage occur at the same time scale as the brittle fracture spalling punching radial fracture ductile penetration petaling penetration at the projectile tip and wake, in addition to damage of reflected waves (e.g., spalling, petaling).

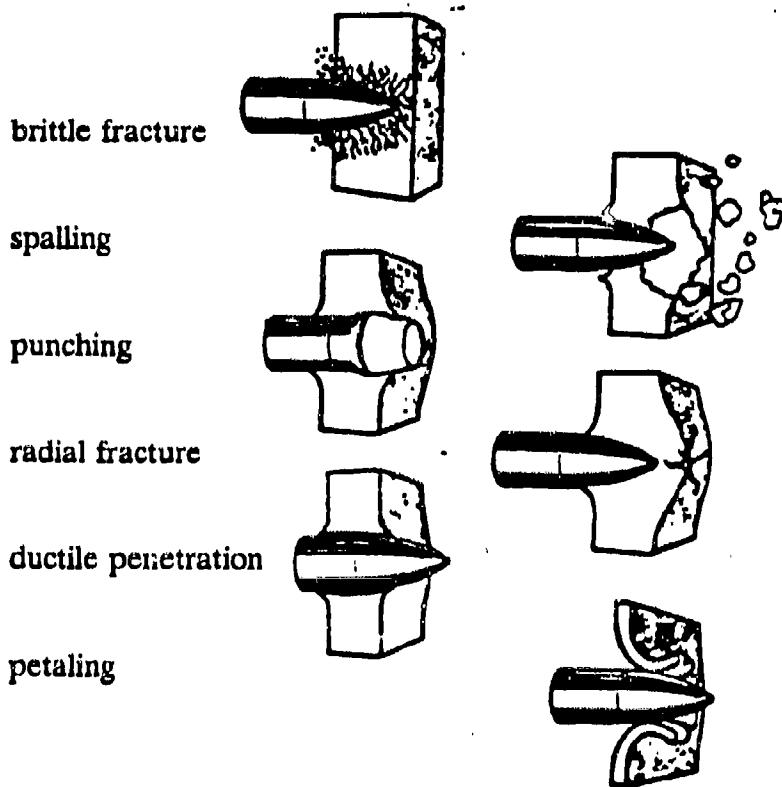


Figure 2. Mechanisms of Failure During KE Penetrator Impact
(Ferrari, 1988)

For low velocity impacts, reflected stress waves may induce rather extensive internal material cracking either in diffuse fashion or along specific planes ahead of the penetration. Brittle target materials such as ceramics are particularly susceptible. Plugging failure is also common for more ductile materials which form adiabatic shear bands through the target plate thickness, resulting in localized shear deformation and punching of a plug of material out of the back side of the target; low strain hardening, high strength metals are particularly susceptible to this form of localization. Plugging failure is common for doubly curved plates, and for blunt projectiles with kinetic energy provided more by mass than by velocity.

Spall failures are common for doubly curved plates with hard armor; this mode of damage is intermediate to bending and punching failure modes. It progresses from the interior of the target plate to the exit surface, resulting in a crater on the exit face and is associated with coalescence of damage reinforced by tensile stress wave reflections. The exiting fragments are lethal to interior systems and occupants. At low projectile velocities, the spall ring may exit as a plug at the same time scale as perforation of the penetrator. For hypervelocity impacts, the ejection of the spall ring or fragments thereof may occur at several wave reflections after the penetrator exit, depending on plate thickness and the longitudinal wavespeed in the target material relative to the impact velocity.

Hypervelocity penetrators, and even small diameter hypersonic penetrators, may pierce armor with accompanying ablation of target material (i.e., erosion), as well as erosion of the penetrator.

Gross structural bending failures may occur when predominately flat, compliantly supported plates are impacted by soft/heavy/blunt nose projectiles at low to moderate velocities (e.g., artillery velocities). This mode of failure may induce significant internal target cracking and perhaps perforation/tearing, but limited fragmentation. Such failures are also typical of blast wave damage outlined earlier.

Fragmentation is another important failure mode, since it may induce additional damage to components on the protected side of the target material. Fragmentation is typically associated with intermediate to high velocity impacts. A distribution of fragment sizes is typically observed which is highly dependent on material microstructure length scales, but remains somewhat poorly understood.

Several relevant material failure mechanisms may be apparent in a post-mortem examination for each penetrator/target material combination, with the relative influence of different failure mechanisms dependent on projectile velocity. As a rule, deformation and damage processes become more highly localized in nature with increasing impact velocity, although the influence of structural geometry and constraints on stress wave reflections tend to cloud this generalization.

Ceramics must be treated differently from metals, as they dissipate energy by microcracking rather than plastic deformation. While for metals the strength of the target correlates with penetration depth and hole size, both the dynamic elastic limit and the fracture toughness are important for ceramics. The impact resistance of ceramic targets depends highly on the degree of target geometry and confinement (cf. Shockley et al., 1990).

1.2 Shaped-Charge Jets

Shaped-charge jets are produced by the detonation of an HE behind a thin conical liner material, typically copper, to produce a small diameter "jet" of material with extremely effective penetrating power through monolithic, passive armors. The HE is detonated by a signal from the nose cone with a stand-off distance on the order of one caliber diameter. The stand-off distance for a shaped-charge jet is extremely important in terms of jet penetration. Too long of standoff results in breakup of the jet and lack of penetration. Too short of standoff results in lack of stretching and jet formation which decreases the tip velocity and penetrating power. The explosively formed liner material stretches into an axisymmetric jet with a peak velocity of approximately 10 km/s at the tip; the velocity at the trailing end of the jet is only on the order of 1 km/s. Roughly 10-20% of the HE energy is transferred to kinetic energy of the jet; hence, these penetrators are sometimes referred to as chemical energy (CE) penetrators. Peak plate pressures are in the range of 100-200 GPa (Walters, 1990). Average temperatures are on the order of 20-50% of the melting temperature with average strain rates on the order of 10^6 to 10^7 s $^{-1}$. Local temperatures and shearing rates in the vicinity of the tip of the jet are higher still. The process is essentially adiabatic in character. A diagram of shaped-charge jet appears in Figure 3. A flash radiograph of the penetration process is presented in Figure 4.

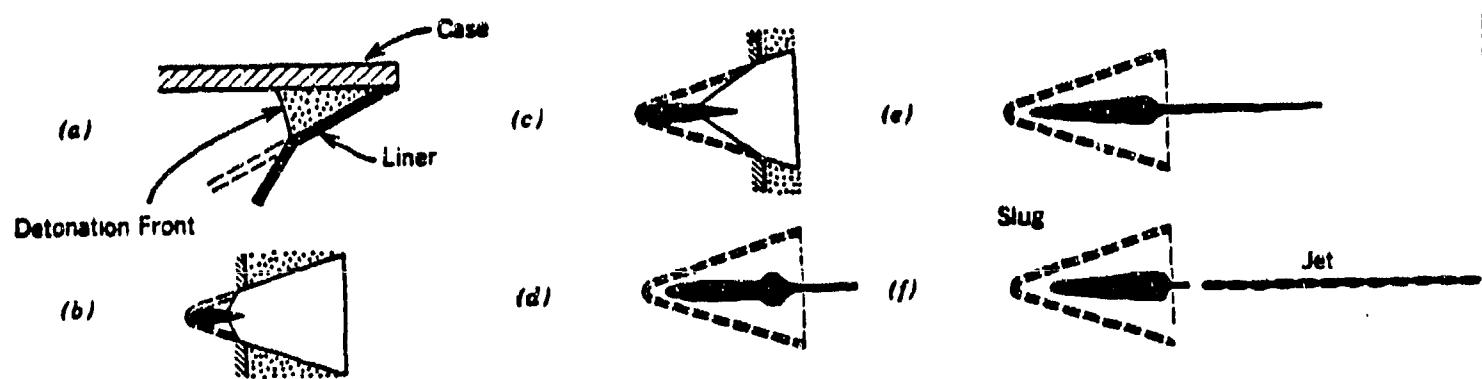


Figure 3. Illustration of the Collapse of a
Shaped-Charge Jet With a Conical Liner
(Walters & Zukas, 1989)

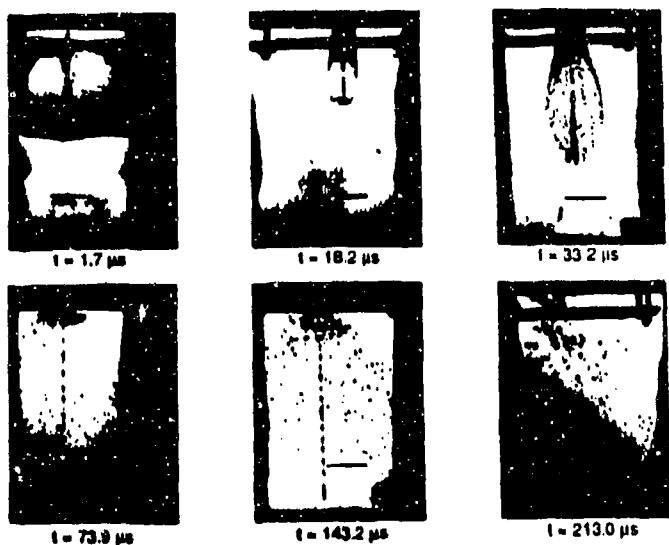


Figure 4. Radiographs Series for OFHC Copper Shaped-Charge Jet Impacting RHA Plate at a Standoff Distance of 12 Charge Diameters (Raftenberg, 1992b)

Target damage is characterized by extreme erosion of material and highly localized damage in the form of adiabatic shear banding, microcracking, solid state phase transformations, and even melting. Fragmentation is commonly observed after the penetration is complete (cf. Figure 4).

Tandem charges can be used to attack composite or reactive armor, but neither tandem nor single shaped-charges typically inflict behind-the-armor damage of widespread proportions.

1.3 Explosively Formed Projectiles (EFPs)

Sometimes called self-forging fragments, EFPs start as parabolic (or other shape) liners, similar to shaped-charge jets, which are explosively formed into a rod-like penetrator traveling at velocities as high as 3-5 km/s (cf. Figure 5). In contrast to shaped-charges, EFPs are effective from stand-off distances as far as a hundred caliber diameters from the target. It is imperative that fracture of the EFP is avoided during forming; FCC materials tend to form adiabatic shear bands which promote fracture. BCC materials such as tantalum may be superior provided they are initially soft and resist formation of shear bands, permitting formability without damage, but strain harden greatly, preventing flattening against the armor (Hornemann et al., 1987).

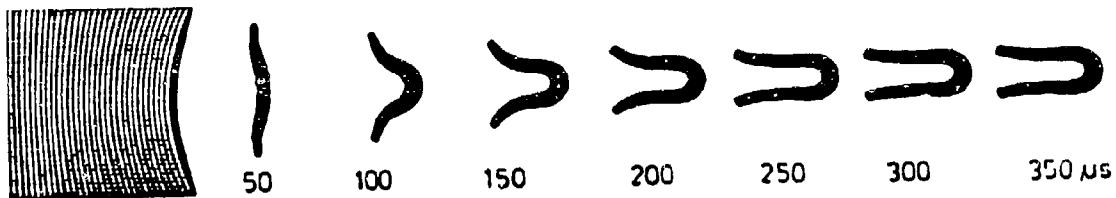


Figure 5. Numerical Simulation of EFP Formation
(Hornemann et al., 1987)

1.4 Other Penetrators

In addition to the KE and CE penetrators, fragmenting munitions may be used for attacking lightly armored targets or personnel. The issues associated with fragmentation of these munitions have much in common with high velocity KE or CE penetrators. A summary of typical pressure, temperature, strain and strain rate conditions for various classes of penetrators appears in Table 2.

Table 2.
(NMAB Report, 1980)

	Pressure GPa	Homologous* Temperature	Strain	Strain Rate s ⁻¹
Gun Launched (0.5–1.5 km/s)	Peak ~ 20–40 Avg. ~ 3–5	Peak ~ 0.2–0.3 Avg. ~ 0.1	Peak >1 Avg. ~ 0.2–0.3	Peak ~ 10^6 – 10^7 Avg. ~ 10^4 – 10^5
Self-Forged Fragment (1.5–3 km/s)	Peak ~ 70 Avg. ~ 10	Peak ~ 0.4–0.5 Avg. ~ 0.2	Peak ~ 1 Avg. ~ 0.2–0.3	Peak ~ 10^6 Avg. ~ 10^4 – 10^5
Shaped-Charge Jet (3–10 km/s)	Peak ~ 100–200 Avg. ~ 10–20	Peak > 1 Avg. ~ 0.2–0.5	Peak >> 1 Avg. ~ 0.1–0.5	Peak ~ 10^6 – 10^7 Avg. ~ 10^4 – 10^5

1.5 Summary

Clearly, the characterization of material response and penetrator performance for the applications considered here must rely on a combined testing and simulation program, given the extremes of dynamic conditions encountered. The NMAB Report on "Materials Response to Ultra-High Loading Rates" issued in 1980 concluded that "an iterative design procedure is recommended to optimize the ordnance design process...intimate collaboration is required between computational, dynamic material testing, and ordnance test firing groups."

Armor/anti-armor design and performance evaluation has relied heavily on the use of hydrocodes for dynamic analyses which capture many salient aspects of the dynamics. To date, these simulations have been used primarily to explain experimentally observed phenomena, and to only mildly extrapolate beyond conditions for which the codes are validated. As pointed out by the 1980 NMAB Report, "the most serious limitation, at present, concerns the modeling of dynamic material failure." This weak link was echoed in the review of Zukas et al. (1981) in their review of hydrocode developments and material modeling. Modeling of dynamic material failure persists as a weakness. Since only the post impact features of target failures are readily measurable, it is extremely difficult to assess the viability of the damage mechanisms assumed in existing hydrocodes. As will be discussed, the hydrocode algorithms may alter the intended role of the constitutive equations for material behavior, rendering the problem of modeling dynamic impact even more difficult and challenging. In the following sections, we discuss material behavior under dynamic conditions, surveying various aspects of deformation and failure. We then discuss the state-of-the-art in hydrocodes used for impact simulation. Finally, we outline a thermodynamical framework which admits various damage mechanisms through the concept of internal state variables, accounting for damage evolution, interaction, dissipation and kinematics of deformation in a consistent manner, admitting micromechanical solutions as they become available.

2. DYNAMIC MATERIAL BEHAVIOR

The behavior of damaged elasto-plastic solids under impact conditions is extremely challenging to investigate experimentally and to model. Historically, equations of state for fluids have been augmented by simple deviatoric incremental plasticity models with either perfect plasticity or isotropic hardening to address the elasto-plastic propagation of stress waves in impacted, plastically deforming metals. With the passage of time, it became clear that it was necessary to include material damage effects to properly capture structural response; accordingly, various algorithms were introduced in hydrocodes to simulate local material failure. Early approaches, still resident in many hydrocodes as options, include a tensile pressure cut-off for material fracture (cf. Hallquist, 1988), the maximum principal stress criterion (cf. Hallquist, 1988), and cumulative strain to failure (cf. Wilkins et al., 1980). Simple criteria such as critical stress are instantaneous in nature; when they are met, there are several further options to consider regarding subsequent material response. In some cases, the resistance to tensile normal stress or pressure is reduced to zero; this condition may be augmented by enforcing zero resistance to shear stress. Such failure models degrade material resistance to deformation instantaneously rather than progressively, and have demonstrated some correlative capability for impact problems. However, developments of the last two decades have increasingly focused on tracing the evolution of various forms of damage through the impact history. Central to any such cumulative damage model is the role of dynamic material deformation response.

2.1 Modern Theory of Elasto-Plastic Shock Waves

The work of Wilkins (1963) laid the foundation of the present approach to elasto-plastic analysis of shocks. In this approach, the plastic deformation of the material is assumed to be associated with the deviatoric stress and is decoupled from the hydrostatic relation which is based on a pressure dependence on density and internal energy per unit of undeformed volume. The equation of state can be selected such that the familiar bulk modulus from linear isotropic elasticity is approached for pressures below the Hugoniot Elastic Limit (HEL). Such generalized fluid models for the hydrostatic case are strictly understood to be an approximation for solids, since they can support shear stress. Moreover, the evolution of material damage which contributes to change of specific volume complicates the estimation of a history-dependent shock Hugoniot curve. From the thermodynamic viewpoint, more sophistication is required to characterize the dependence of the free energy function on the evolving damage.

Key aspects of the deviatoric material response (i.e., shear) at high strain rates include pronounced strain rate- and temperature-dependence of yielding and plastic flow, and strain hardening/softening behavior. Although the yield strength may be small compared to the pressure of the shock wave for high velocity impacts, the deviatoric plastic flow of the material is very important; plastic flow modifies the elastic shock wave by introducing high amplitude plastic waves which trail the elastic wavefront, significantly altering the shock wave profile and the associated shock impulse (cf. Wilkins, 1963). Moreover, the shock wave profile is dispersed and smoothed by material strain rate-dependence, strain hardening and Bauschinger effects (NMAB report, 1980). Other effects, such as material damage or solid state phase transformations, may also significantly alter the shock wave profile.

It is within the preceding general context that material models have developed for shock problems. The great majority of simulations have been performed under the assumption of material isotropy, even when damage parameters have been introduced. Other assumptions are difficult to fully justify in view of the uniaxial character of shock experiments such as split Hopkinson bar or flyer plate impact experiments which characterize homogeneous uniaxial shock response of materials at strain rates on the order of 10^2 s^{-1} - 10^4 s^{-1} and 10^4 s^{-1} - 10^6 s^{-1} , respectively. While actual impact scenarios involve multiaxial straining due to irregular geometries and mode of impact, it is difficult to assess the dynamic material response under such strain states; only limited data are available for such cases (e.g., oblique flyer plate impacts which generate both shear and longitudinal waves, exploding constrained cylinders, torsional Hopkinson bar, double shear Charpy and steep projectile impact (Mescall & Rogers, 1987)). In addition, information regarding material response to multiple shocks is valuable in view of the multiple wave reflections experienced in actual impacts. Phenomena such as shock hardening and damage evolution require multiple shocks for clarification, based on the successive HEL and Hugoniot curves obtained. Effects of damage anisotropy require either consideration of reflected waves or multiple shocks in orthogonal directions, i.e., sequence experiments. Due to the complexity of such experiments, these phenomena are commonly implicitly included in parameters of constitutive laws such as the dynamic yield strength and the equation of state for pressure dependence. Material viscosity, strain hardening and Bauschinger effects tend to smear elastic-plastic shock waves, often rendering direct interpretation difficult. The details of the progressive failure process in ceramics lead to similar considerations (cf. Mandell, 1993).

A widely employed procedure, recommended by the 1980 NMAB report, is to adopt an elastic-perfectly plastic, rate-independent plasticity model with, at most, a temperature-dependent dynamic yield strength. A shortcoming of this approach is that the selection of the dynamic yield strength is often made on the basis of agreement with a numerical impact solution for a particular case. It may also be estimated on the basis of the HEL determined from flyer plate or split Hopkinson bar experiments, which involve additional assumptions for interpretation. The use of perfect plasticity with a dynamic yield strength may be most appropriate for high strength metals with low strain rate sensitivity for impact conditions which involve moderate temperature rise; such materials tend to localize deformation readily in the form of adiabatic shear bands. For many metals which strain harden significantly, exhibit strain rate sensitivity and temperature-dependence, the perfectly plastic description can be a gross simplification. While the prediction of the overall character of impact events, EFP formation, and similar dynamic problems may be reflected by simulations based on a dynamic yield strength and perfect-plasticity, evolution of material damage resulting in material fracture, spallation and fragmentation is not well-correlated, in general, with damage parameters such as cumulative effective plastic strain.

2.2 "Plastic" Equations of State

The next level of sophistication is the assumption of a "plastic equation of state" as proposed by Johnson and Cook (1983, 1985), i.e.,

$$f = \frac{3}{2} s_{ij} s_{ij} - R^2 = \bar{\sigma}^2 - R^2 , \quad \dot{\epsilon}_{ij}^p = \lambda \frac{\partial f}{\partial \sigma_{ij}} \quad (5)$$

where f is the dynamic yield condition, and σ_{ij} and s_{ij} are the components of Cauchy stress and deviatoric stress, respectively; R is the radius of the von Mises yield surface, given by

$$R = [A + B(\dot{\epsilon}^p)^N] \left[1 + C \ln \left(\frac{\dot{\epsilon}^p}{1.0 \text{ s}^{-1}} \right) \right] \left[1 - \left(\frac{T - T_r}{T_m - T_r} \right)^M \right] \quad (6)$$

where T_m and T_r are the absolute melting temperature and room temperature, respectively. The scalar multiplier λ is determined by the consistency condition that the stress state must remain on the rate- and temperature-dependent yield surface during plastic flow. The equivalent plastic strain and strain rate in equation (6) are given by

$$\dot{\bar{\epsilon}}^p = \left(\frac{2}{3} D_{ij}^p D_{ij}^p \right)^{1/2}, \quad \bar{\epsilon}^p = \int_0^t \dot{\bar{\epsilon}}^p dt \quad (7)$$

where D_{ij}^p is the plastic rate of deformation tensor. In the Johnson-Cook failure model, material separation is assumed to occur when the damage parameter, D , reaches unity, where D is defined by

$$D = \sum \frac{\Delta \bar{\epsilon}^p}{\epsilon_f} \quad (8)$$

Hence, this failure model is equivalent to exhaustion of ductility. Stress state and temperature effects on the strain to fracture, ϵ_f , are expressed by the empirical relation

$$\epsilon_f = [D_1 + D_2 \exp(D_3 \sigma^*)] \left[1 + D_4 \ln \left(\frac{\dot{\bar{\epsilon}}^p}{1.08^{-1}} \right) \right] \left[1 + D_5 \frac{T - T_r}{T_m - T_r} \right] \quad (9)$$

where D_1, \dots, D_5 are constants and the triaxiality factor is given by

$$\sigma^* = \frac{P}{\sigma} \quad (10)$$

Here, P is the pressure.

This isotropic hardening plasticity model characterizes strain hardening, temperature-dependence of flow stress, and instantaneous strain rate sensitivity. While this approach represents an improvement over perfect plasticity and contains several salient features, it is a highly idealized representation of evolutionary aspects of material hardening and deformation which are manifested under complex loading and unloading (reverse yielding), nonproportional loading, and non-isothermal conditions characteristic of blast waves and projectile impacts. Under sequential, nonproportional loading conditions which are generally associated with reflection of stress waves, a law of this nature must be considered as highly approximate. Nonetheless, this type of model enjoys widespread usage due to its simplicity and availability of parameters for various materials of interest (Johnson & Holmquist, 1989).

The Johnson-Cook model and other similar rate-dependent plasticity models may be categorized as equations of state based on "observable" variables such as plastic strain, plastic strain rate and temperature. There are several grounds on which to fundamentally question this class of models. First, they are ill-suited to incorporate detailed representations of progressive deformation and failure mechanisms, and likely have very limited capability of extrapolation to high strain rates. For example, it may be granted that diffusion effects do not have time to operate effectively given the time scales of the impact problem. It is well-known that there are at least three forms of softening behavior which may occur at high homologous temperature and large strains under quasi-static loading conditions (i.e., strain rate $< 1 \text{ s}^{-1}$ for sake of argument), including diffusion-controlled static thermal recovery or annealing, dynamic recovery (e.g., recrystallization) and shear localization, which may be promoted by porosity. Suppose that at a much higher strain rate, the same local temperature is reached as in a quasi-static experiment in which temperature is controlled; in this case, static thermal recovery will be inoperative and, because of a link between the threshold for plastic flow, the strain-hardening rate and strain rate (e.g., shock hardening), softening may be induced predominately by formation of adiabatic shear bands, perhaps acting in concert with voids and void sheets. Clearly, the inability to depict the change of the failure mechanism in the observable cumulative strain/strain rate/temperature state space reveals a basic deficiency of this class of models. There are also cases (e.g., low velocity impacts of thin shells, blast waves) where static thermal recovery effects may not be negligible.

Another deficiency of this class of models is the representation of thermal and strain rate history effects which prevail in impact problems. For example, the material ahead of a penetrator undergoes straining in the presence of steadily increasing temperature and strain rate. It is well-known that even under isothermal conditions, metals do not asymptotically follow the stress-strain curve for a given strain rate following a jump from another strain rate after significant levels of plastic deformation. Moreover, for a given strain rate, the subsequent stress-strain behavior at room temperature following significant deformation at a high temperature will not typically correspond to that of an isothermal test at room temperature, and vice versa. In fact, it may not even be close.

In view of this extrapolative deficiency, constitutive laws such as the Johnson-Cook model may essentially be viewed as an alternative, compact method to present experimental results. It is therefore not surprising that fits have been catalogued (Johnson & Holmquist, 1989) for a wide range of materials. They may be expected to perform

reasonably well when characterized at strain rates on the order of the impact problem of interest using split Hopkinson bar tests or flyer plate experiments, apart from certain other features which are lacking.

It is generally recognized that damage couples directly with the viscoplastic relation (and elasticity) of the material, but this coupling is not included in such models; consequently, an instability in the stress-strain response is solely attributed to thermally induced softening due to local temperature increases at high strain rates. In reality, void nucleation and growth form the basis of such instabilities at quasi-static strain rates where temperatures are essentially uniform, and may be expected to contribute significantly to localization of the deformation at high strain rates as well. This lack of coupling exacerbates the problem of judging the suitability of a material model based on hydrocode predictions/correlations. Specifically, as outlined in Section 3, the various failure criteria available in current hydrocodes are essentially used as adjustable parameters to correlate ballistic experiments. Essentially, a coupling is introduced between the stress-strain response and damage only at the scale of the structure rather than the local scale since failed material elements are stiffness-degraded in the hydrocode simulations. Although this approach is consistent with the recommendations of the 1980 NMAB report in support of an iterative simulation-testing philosophy, the resulting pseudo-coupled theory is unsatisfactory from both thermodynamic and physical perspectives. Furthermore, there is little hope of embedding local, micromechanically based deformation and failure criteria (another major NMAB recommendation for long term development) in such an approach.

2.3 Internal State Variable Concepts

Over the past 15 years, considerable advances have been made in developing internal state variable models which have the advantages of (i) describing the evolutionary nature of material deformation, including the coupling of rate- and temperature-dependence with material hardening, and (ii) naturally incorporating information regarding the state of damage and its coupling with the process of deformation. Moreover, couplings of strain rate-dependence with strain hardening/softening behavior and high strain rate phenomena such as shock hardening (Meyers & Murr, 1981; Follansbee & Kocks, 1988; Follansbee, 1988) can be accounted for by introducing evolutionary internal state variables for each salient deformation mechanism.

A distinguishing feature of these internal state variable models is the appearance of internal variables representative of the internal state of the material at each stage of the

deformation process. This material state, including dislocation density and configuration, crystalline microstructure, effects of cavities, microcracks, etc. is not readily observable, in general, but controls much of the history dependence of the material behavior. It is straightforward to extend the thermodynamical concept of a sequence of accompanying, constrained equilibrium states to include these internal variables as was proposed by numerous researchers (cf. Rice, 1971; Germain et al., 1983; Lemaitre & Chaboche, 1985; Allen, 1991). Within this conceptual framework, numerous internal state variable models have been developed and implemented (cf. Walker, 1981; Bammann, 1984, 1985, 1990; Miller, 1987; Follansbee & Kocks, 1988; Chaboche, 1989; McDowell, 1992). Variations of the model of Bodner and Partom (1975) have been consistently employed in impact and other classes of high strain rate problems (cf. Bammann, 1984, 1985, 1990; Batra et al., 1993). Since each of the internal variables represents distinct features of the microstructure and the mechanical response, attention may be devoted to the details of the evolution of the internal variables. At one end of the spectrum, evolution laws can be proposed to simply fit data in a manner analogous to the Johnson-Cook model. At the other end, the structure of this approach permits introduction of known solutions of microscale behavior into the macroscale relations (cf. Miller, 1987; Bammann, 1984, 1985; Aifantis, 1986, 1987; McDowell & Moosbrugger, 1992), almost without limit. In some cases, a breakdown of applicability of continuum mechanics or continuum thermodynamics may occur, i.e., detailed modeling of microscale phenomena on the order of a small number of interatomic spacings and time scales on the order of periodic atomic vibrations. In such cases, quantum mechanical or lattice dynamical calculations may be necessary. Relevant length and time scales in impact problems do not typically broach these limits of applicability.

Internal state variable models have become much more prevalent and well-characterized since the NMAB report of 1980, which offered only passing reference to the immaturity of the field. Within the context of internal state variable theory, it is possible to include a wide range of phenomena, including inelastic deformation, phase transformations, void nucleation and growth, microcrack formation, etc. In the local theory, the effect of heterogeneity on the order of microstructural size scales is assumed to be smeared or averaged as regards its influence on deformation and damage evolution. Macroscale phenomena such as spall cracks, shear bands or fragments are regarded as manifestations of local processes which have coalesced. From a practical perspective, the inherent discretization of the numerical solution may impart some nonlocal character to the solution by averaging over a minimum element dimension which is larger than the scale of high gradients of stress, strain and/or damage. The length scale of the mesh in the

numerical solution essentially provides a minimum length scale which is absent in the local approach. Presumably, this length scale corresponds to some physically based quantity (cf. Zbib, 1992), but this is usually not the case since general guidelines do not exist and the material length scales are often unreasonably small to be considered in numerical solutions. Examples include numerical simulation of progressive thinning of shear bands or fragment size distributions. Nevertheless, theories with continuously evolving damage are likely preferable to the use of instantaneous failure criteria in hydrocodes to obtain an ad hoc coupling of deformation and damage at the structural size scale since the former route holds open the possibility of truly predictive theory and simulation. Continued development of nonlocal internal state variable theories may ultimately offer the capability to take local heterogeneity into account to predict phenomena such as shear band thickness, fragmentation, etc., without relying on the numerical discretization or averaging schemes to provide the appropriate length scales (cf. Aifantis, 1986, 1987; Zbib, 1992). Alternatively, inclusion of length scales in the parametric dependence of the free energy and dissipation potentials of thermodynamics based on internal variables may offer similar capabilities. In fact, shear banding and fragmentation remain as two rather unsatisfactorily developed areas primarily due to the length scale and nonuniform distribution issues in modeling, as well as to the statistical nature of these processes. It is likely that phase transformations and perhaps some refinements of void growth and micromechanics are subject to similar considerations.

2.4 Microstructural Effects and Internal State Variables

With this background on internal state variable theories, the remainder of this section will concentrate on the role of microstructural events in high strain rate behavior. A later section will present more details on a thermodynamic framework for internal state variable models. We shall limit our discussion to the microstructural issues controlling the ballistic performance of metallic materials. This discussion will be illustrative of the kinds of issues surrounding microscale phenomena in other materials as well. Ceramics have received much recent attention, and have different but analogous microscale behaviors, e.g., microcracking, comminution and hydrostatic bulking (cf. Rajendran & Cook, 1988; Shockley et al., 1990; Klopp & Shockley, 1992).

There is a strong relationship between the microstructure of a material and its ballistic performance. The microstructure influences not only the deformation behavior of the material, but also the mechanism of failure. This is very well illustrated by the results of Mescall and Rogers (1987), as shown in Figure 6. Here, a normalized ballistic limit is

plotted as a function of the Brinell hardness of 4340 steel. The different hardnesses were obtained by changing the heat treatment of the material, which in turn alters the microstructure of the steel. Also shown is the performance of the same 4340 steel processed by vacuum induction melting (VIM) and electroslag remelting (ESR). At low hardnesses the steels fail by plastic flow; as the hardness of the steel increases, failure occurs by the formation of shear bands. The formation of shear bands leads to failure by plugging. At still higher hardnesses the steel fails by disking wherein the steel delaminates near the rear of the target along planes parallel to the impact face. However, simply changing the manner in which the steel is processed leads to a dramatic change in the ballistic performance of the steel, since the ballistic performance of the ESR steel is considerably superior to that of the VIM steel. As all of these effects can be ascribed to different microstructures, the importance of microstructure in the modeling of the ballistic performance of a material is clear, particularly in the hypersonic velocity range.

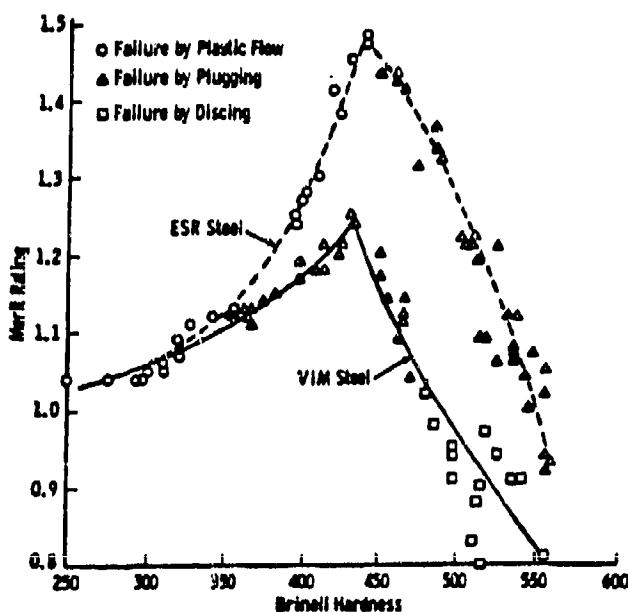


Figure 6. Ballistic Merit Rating Versus Hardness for VIM and ESR Steels
 Precipitous drop in performance is due to onset of adiabatic shear banding
 (Mescall & Rogers, 1987)

2.4.1 Dislocation Substructures

Dislocation substructures which form during the compressive shock (cf. Meyers & Murr, 1981) in one-dimensional Hopkinson bar and flyer plate experiments are found to be much more homogeneous and uniform in character than those which arise during quasistatic loading at much lower strain rates; these structures may in fact be unstable under continued deformation. There is some evidence that the deviatoric stress required for plastic flow increases significantly at pressures up to 100 GPa (cf. Chhabildas & Asay, 1978), a manifestation of shock hardening. Such observations suggest evolutionary laws for material hardening which are dependent on strain rate rather than just a strain rate dependence of flow stress (cf. Follansbee & Kocks, 1988; Mecking & Kocks, 1981; Follansbee, 1988; McDowell & Moosbrugger, 1990). In other words, the rate of strain hardening depends on strain rate. It is very difficult to fully assess the influence of shock hardening on dynamic response; quasi-static experiments following the shock are not fully satisfactory in view of the strong tendency towards re-organization of dislocation substructures into more stable configurations. Inference of shock hardening effects from the Hugoniot curve or the shock wave profile is also somewhat ambiguous. It is likely that internal state variable theories which are able to reflect hardening effects associated with both the homogeneity of the dislocation distribution under shock loading and quasi-static low energy dislocation substructures such as cells (cf. Bay et al., 1992; Sidoroff & Teodosiu, 1986; Teodosiu, 1991; Miller & McDowell, 1992) may offer considerably more promise than observable state variable theories across a wide range of loading rates. There is also a marked coupling of the elastic constants, including both the shear and bulk moduli, with pressure (Meyers & Murr, 1981) due to the extreme degree of lattice compression; this effect requires a consistent treatment in terms of the free energy.

Effects of shock pressure on both deviatoric and hydrostatic response may be most fundamentally addressed using lattice dynamical calculations. For example, Monte Carlo simulations were performed (Taylor & Dodson, 1985) for 2-D lattices composed of approximately 300 atoms for an HCP structure subjected to slow and fast (relative to lattice equilibration time) rates of uniaxial compression and hydrostatic compression. The Lennard-Jones interatomic potential was used and the lattice assembly was subjected to periodic boundary conditions. It was found that (i) the higher rate (e.g., shock) uniaxial loading incurred significantly more lattice damage (point defects) than the slow loading, and (ii) hydrostatic loading incurred no damage for either fast or slow loading rates. Similar calculations with initial conditions of lattice damage would shed light on the

appropriate equation of state for the shocked material in the presence of realistic forms of defects such as line defects and microcracks. At the present time, however, our computational capability is too limited to address this level of detail. Apart from consideration of computational practicality, there is an issue of which interatomic potential is most appropriate or realistic.

2.4.2 Formation of Shear Bands

Strain localization, or shear band formation, appears to be the prevalent failure mechanism for steels above a certain hardness (cf. Figure 6). The presence of shear bands is clearly observed in metallographic sections which reveal highly deformed martensite laths which show alignment along the direction of flow as well as the appearance of white etched bands. The white etched bands are typically smaller in size and reside within the larger bands of deformation. Many researchers have noted the appearance of the white "transformed" band and concluded that the temperature of the center of the bands has exceeded the austenitizing temperature. Although the measurement of the extreme hardness of these white bands is consistent with austenite being rapidly quenched to form martensite (Rogers and Shastry, 1980), recent transmission electron microscopy appears to indicate that there is no major difference between the microstructure within these bands and that of the undeformed alloy (Wittman et al., 1990). Instead, it is proposed that the carbides dissolved on heating are re-precipitated as smaller carbides on dislocations. The carbides pin the dislocations and thus increase the hardness within the band. However, Wittman et al. (1990) propose that the carbon diffusion and re-precipitation of carbides is strictly due to the elevated temperature within the bands and ignore the possibility that the high level of stress within the band could, in itself, lead to carbide dissolution. Thus, deducing the thermal history of the band by measurements of the final microstructure can be misleading. Nevertheless, when modeling the flow properties of the band, one should assume that the microstructure is martensitic. The importance of these observations is that it is essential to understand the microstructure within a band to model the flow properties of the shear band. This premise has been largely ignored to date in modeling shear bands which form at high strain rates.

Simple analytical models of shear band formation indicate that the temperature-dependent strain hardening rate and strain rate-dependence are two features that can be used to either delay or stabilize a region undergoing shear localization (cf. Olson et al., 1980). Unfortunately, neither of these two features appear to be particularly amenable to change by alterations of the microstructure. Thus, it appears advantageous to address the factors

controlling the nucleation of the shear bands. Once these factors are identified, adequate models can be developed and incorporated into existing codes.

Most researchers have studied the onset of shear band formation as a result of a thermally driven mechanical instability (cf. Wright & Batra, 1985; Wright & Walter, 1987; Lipkin et al., 1988; Molinari & Clifton, 1987). There are many possible heterogeneities which can trigger the formation of a band in a region of material which is on the verge of instability. For example, there appears to be growing evidence that microvoid formation may be a precursor to shear band formation. The experiments of Cowie et al. (1987) reveal a clear pressure dependence of the instability shear stress, as was pointed out earlier by Olson et al., (1980). In fact, Cowie et al. (1987) found the shear instability strain in 4340 steel to be essentially the same at both quasi-static and intermediate loading rates, in spite of quite different temperatures and flow stress levels, suggesting that microstructural instability dominated rather than thermal softening. They present microstructural evidence showing the presence of $0.5 \mu\text{m}$ voids around a pair of undissolved alloy carbides. Their experiments also indicate that the fewer secondary particles present in the structure which may act as nucleation sites, the higher the instability shear strain. The possible relationship between void formation and the development of shear bands is evident in the microstructures shown by Krause and Raftenberg (1993) in RHA plates which have been perforated by shaped-charge jets. In their study, it is clear that there are large voids present in the shear band ahead of a crack. Finally, these experimental observations of microvoid nucleation of shear bands are consistent with the analysis of Hutchinson and Tvergaard (1987). Thus, it may be necessary to include the factors leading to the formation of microvoids in modeling the development of shear bands.

Use of the Zener-Hollomon parameter (cf. Zener & Hollomon, 1944) implies a thermal softening mechanism which may only partially contribute to the localization. Mescall and Rogers (1987) argue that the consequences of any nucleated microvoids on shear localization in a zone of very high hydrostatic compression must be minimal; moreover, the pressure-dependence of the instability strain reported by Cowie et al. (1987) implies that microvoid nucleation (and certainly growth) would be significantly suppressed at hydrostatic pressures encountered in ballistic events. However, these conclusions were based upon considering the macroscopic stress state of the body. If the material is inhomogeneous, e.g., with distributed precipitates, grain boundaries, etc., the stress state in the local neighborhood of one of the heterogeneities may be strongly deviatoric. It is likely, therefore, that thermal softening indeed plays a strong role in nucleating shear

bands, but microvoid nucleation introduces material length scales for instabilities that couple with rate-dependence to establish instability strains and shear band thicknesses.

There is compelling evidence that shear bands form along preferred macroscopic directions, a source of anisotropy. Shockley and Erlich (1980) investigated shear band formation in 4340 steel at various levels of hardness. In particular, they investigated the effects of the rolling direction of the plate on the nature of the failure. For moderately high loads, a strong orientation dependence or fibering of the failure was observed; the failure was observed to be parallel to the rolling direction. This is in contrast to the tensile properties, which were independent of the rolling direction. The cause of the anisotropy in the failure mechanism is thought to be inclusions which become elongated into stringers during the rolling process. At higher loads, however, the orientation dependence disappeared. This implies that these anisotropy effects are crucial in understanding spallation and fragmentation processes, since these processes may occur at comparatively low loading conditions, and points to the need for developing anisotropic damage models (cf. Seaman et al., 1985). There is also evidence that shear banding may be most prevalent in weakly strain hardening FCC metals, and that BCC structures may exhibit delayed shear localization by comparison.

2.4.3 Texture, Substructure and Anisotropy

In addition to shear banding, there is another source of anisotropy which is almost always overlooked. Anisotropy of the material deformation behavior in shear preferential to compression, typically associated with texturing effects, is a first order effect common to all metals at large strains that is not included in most of the "observable" state variable models which are based on J_2 plasticity. This shortcoming may bear significantly on penetration problems where essentially compression experiments are used to fit models for stress-strain behavior, but the deformation in the vicinity of the penetrator is largely of shear character. Bay et al. (1989), Hughes and Nix (1989), Hughes and Hansen (1991), Teodosiu (1991), etc. have shown that dislocation substructures and geometrically necessary boundaries or microbands form differently in shear than in compression, contributing almost equally to texture in producing the disparity in effective flow stress between compression and torsion at the same effective plastic strain level (Rollett et al., 1992). Miller and McDowell (1992) have related this disparity to the third invariant of deviatoric stress, following the arguments of Drucker (1949), in a form suitable for modification of numerous internal state variable models. It is also important to include Bauschinger effects, in general, in macroscale plasticity relations suitable for general

loading paths, an element which is absent from isotropic hardening theories. Huntington (1991) has undertaken a similar route.

It should be pointed out that crystal plasticity theory, a continuum approach which explicitly accounts for slip planes, slip directions and intergranular constraints, is presently incapable of modeling the effects of such dislocation substructure without further modification, although texture is readily predicted. On the other hand, texture development is very difficult to include in macroscopic continuum theories. One logical possibility is to treat texture as an internal state variable in its own right over some representative volume of material. Twinning is a potentially very significant mode of deformation at high strain rates which is often overlooked. The decomposition of the material volume domain into twinned regions and regions undergoing crystallographic slip is very relevant to details of material damage evolution, analogous to the case of solid state phase transformations.

The directions along which RHA fails can be strongly anisotropic. This is due, in part, to the solidification process used to produce the steel. In this case, interdendritic segregation results in a strongly nonuniform solute distribution within the material which cannot be removed by subsequent thermomechanical processing. Upon rolling, these regions of chemical inhomogeneity become elongated along the rolling direction, leading to anisotropic material properties. There have been many studies showing that such nonuniform solute distribution exist within RHA. The work of Olson et al. (1983) showed that these composition striations can lead to an anisotropic embrittlement process wherein failure occurs in the elongated regions of high solute content. This is consistent with the fibering studies of Shockley and Erlich (1980) and the observation that the solidification processing technique (ESR versus VIM) can have a major impact on the penetration resistance of the steel. It may be concluded from this work is that the failure properties of armor can be strongly anisotropic simply due to the solidification processing technique used to produce the armor. Models which neglect this anisotropy may be overlooking an important element controlling the ballistic penetration resistance of steels.

A very difficult issue lies in relating quasi-static multiaxial behavior to dynamic plasticity. To this end, combined stress state dynamic experiments would be very useful. Unfortunately, standard split Hopkinson bar and plate impact experiments are intrinsically one-dimensional in strain. Techniques have been developed for plate impact and for Hopkinson bar tests to introduce both shear and compression-tension. However, the isolation of these stress states and their effect on deformation is nontrivial.

2.4.4 Void Coalescence

At high strain rates, many materials fail by the coalescence of microvoids into void sheets. While the factors controlling the growth rate of these microvoids are reasonably well understood, the phenomena controlling the nucleation of these voids is not. Although void growth is suppressed under compressive loading typical of shocks, they can nucleate profusely. At a later time, when reflected stress waves induce tensile stresses, these nucleated voids can then readily grow. Voids can be nucleated, in principle, at any inhomogeneity. However, there is much evidence that microvoid nucleation in RHA begins at particle-matrix boundaries (Senior et al., 1986; Cowie et al., 1987). With this in mind, Needleman determined the factors controlling the nucleation of microvoids at a particle-matrix interface. He found that the primary factors controlling the nucleation process are the strength of the bond between the particle and the matrix and the spacing between particles. The model of Rice and Johnson (1970) also indicates a strong particle spacing dependence on fracture via void coalescence. This has an important impact on the modeling of microvoid nucleation, as it implies that the nucleation rate is a function of both alloy chemistry and heat treatment. The experimental work of Garrison et al. (1987) indicates the importance of maximizing the spacing between inclusions to enhance the fracture toughness of high strength steels. From a modeling standpoint, once the precipitates in an alloy have been identified, it is possible to use a continuum void nucleation model, along with a quantum mechanical calculation of the strength of the interface, to produce a 'first principles' model of the nucleation process which can be used in numerical simulations. These results also indicate that an accurate model of the void nucleation process will have to consider the spatial distribution of second phase particles which can nucleate voids.

2.4.5 Phase Transformations

Most simulations of the armor penetration process neglect the effects of phase transformations on the constitutive relations of the material. The phase transformation processes which affect the properties of armor steels range from the solidification process used to produce the armor to those which are induced by the penetration process itself.

The stresses generated by the propagation of shock waves through a material can be quite large, frequently orders of magnitude higher than the deviatoric flow stress. The magnitude of these stresses is sufficient to induce solid state phase transformations. The majority of finite element calculations of penetration processes neglect such solid state

phase transformations. However, a few calculations have shown that such transformations can have a marked effect on the development of damage in armor. For example, Shockey et al. (1975) found it impossible to predict the correct spalling process without the inclusion of the pressure induced phase transformation in iron. While these results are interesting, RHA is not pure iron; at high hardness levels, it consists primarily of tempered martensite. Thus, the implications of this calculation are unclear. However, there have been a number of studies which indicate that the extreme stresses accompanying a shock can induce solid state transformations in RHA. Rhode (1970) showed that a shock wave can cause martensite to transform to austenite. The effects of such a phase change can be significant since the changes in the local density and elastic properties of the material will alter the manner in which elastic waves propagate in the material. As these shock induced transformations are known to occur, it is clear that such changes should be included in existing hydrocodes.

The temperature of the material in the region close to the penetrator-armor interface can be quite high. However, the possibility of melting of either the projectile or armor has not been considered in detail. While the temperature of the material is frequently calculated in many finite element codes, the temperatures determined in such codes may be below the melting temperature of the material due to the coarse element or cell size used in the calculations compared to the small region directly near the penetrator-armor interface. There is good experimental evidence that melting occurs during the penetration and fragmentation process. Metallographic studies of RHA fragments by Krause and Raftenberg () shows a clear dendritic microstructure along the outside of the fragment. As this dendritic structure is not present in the RHA, the presence of dendrites implies that the external surface of the fragment was liquid during the fragmentation process. In addition, metallographic studies of KE penetrators by Day and Hogan (private communication, August 1993) have shown clear evidence of bubble and eutectic formation along the sides of the hole produced by the penetrator. Both of these observations are consistent with the presence of a liquid phase during penetration. The impact of these observations on the constitutive laws employed in finite element codes are potentially profound.

2.5 Spallation/Fragmentation/Erosion Mechanisms and Models

Spallation refers to the fracture of a material plane internal to the impacted surface, possibly followed by ejection of material. Spalling typically occurs in conjunction with the development of tensile stress waves arising from reflections from the back surface of the target plate. While the most popular current spall failure model is based on a critical

instantaneous tensile stress or pressure (cf. Hallquist, 1988), even for ductile metals, it is known that this behavior depends in an evolutionary sense on the history of plastic strain and damage. For example, Bertholf et al. (1975) showed that for both Lagrangian and Eulerian hydrocode simulations for a nylon sphere impacting an RHA steel plate at 5.2 km/s that the degree and character of spalling observed experimentally was much more accurately predicted when a 13 GPa polymorphic (α (bcc) to ϵ (hcp)) phase change was included in the target material. The character of the stress waves were altered substantially by the phase change. Possible influences on spall fracture which have not been widely considered include void nucleation and growth, solid state phase transformations, material anisotropy associated with rolled target plate material. Tuler and Butcher (1968) were the first to propose an evolutionary form for spall damage with an hereditary integral based on an empirical relation between stress and spall fracture strength, including impulse stress and duration.

Fragmentation remains one of the most difficult and elusive failure mechanisms. The key difficulty lies in characterizing the statistical populations of fragment size and energy, and to predict/correlate this behavior. It is one of the most relevant failure mechanisms in terms of vulnerability/lethality analysis since fragmentation accounts for much of the auxiliary damage incurred by impact of projectiles or shaped-charge jets (cf. Trucano, 1993). It is generally recognized that the fragment size distribution depends on the projectile kinetic energy and velocity, with a finer distribution of fragments with increasing velocity (cf. Backman & Finnegan, 1985; Quidot, 1987). Grady (1990) has set forth a quite general argument, substantiated by data (cf. Mock & Holt, 1983), that the fragment number distribution for a given impact is an exponentially decreasing function of fragment mass. Several attempts have been made at modeling fragmentation, including the models of Grady and Kipp (1980, 1982, 1985, 1989), the statistical model of Dienes (1984), the model of Quidot (1987) and recent work by Yuan (1990). Most of these models seek to balance the projectile kinetic energy in some manner with the creation of new surfaces during fragmentation using the dynamic analog of Griffith fracture theory. The formation of debris clouds moving with the penetrator in hypervelocity impact typically precede the occurrence of fragmentation. Hence, both spallation and fragmentation are complex phenomena which depend on wave reflections and the history of material deformation and damage.

The NAG-FRAG models developed at SRI by Seaman et al. (1985) consider the nucleation, propagation and coalescence of microcracks in order to more accurately

describe both spallation and fragmentation processes. As such, these are among the most comprehensive continuum models which have been applied to the impact damage problem. A key element is the coupling of the damage with the compliance of the degraded material. Curran et al. (1977) list six ways in which a deformation at an arbitrary rate can be accommodated: (1) elastic distortion, (2) homogeneous plastic flow, (3) phase changes, (4) nucleation and growth of ductile microvoids, (5) nucleation and growth of brittle microcracks and (6) nucleation and growth of shear instabilities. Hence, polymorphic phase transformations, microvoid nucleation and growth, shear banding and microcrack nucleation/growth are all relevant to spallation/fragmentation processes within this framework. Moreover, they all potentially contribute to material softening which assists penetration. Spallation occurs by coalescence of microfractures and fragmentation may depend on the linkage of cracks associated with distributed shear bands. The phase change strongly affects the rarefaction wave and changes the character of the shock impulse (cf. Shockley et al., 1975). Curran et al. (1977) and Erlich et al. (1980) at SRI have also proposed a model for distributed adiabatic shear bands (SNAG) which assumes a critical strain and strain rate for nucleation of shear bands. Seaman et al. (1985) have focused further on a continuum model for microfracture and coalescence leading to fragmentation, with promising results. The key to predicting the fragmentation and spallation processes in this approach is proper representation of the distribution of microcracks. These SRI models have enjoyed a measure of success in predicting material damage, including spallation and fragmentation; however, the same forms are employed for nucleation and growth of both voids and microcracks which requires justification. The void growth law corresponds to the growth of a void in a viscous fluid, which perhaps does not fully reflect the situation in a strain-hardening, rate-dependent solid.

Another poorly understood material deformation and failure mechanism is the backward extrusion and ablation/erosion of material during penetrator ingress. This feature is employed in hydrocodes, as discussed in Section 3, to properly match experimental observations of hole size and residual penetrator velocity. However, perhaps due to the difficulty of quantitative assessment of failure mechanisms, there is no well-established basis for the primitive models based on ductility exhaustion which are used in the hydrocodes as the criterion for material erosion. The threshold value for failure has a much greater impact on impact erosion resistance, in general, than does the material yield strength.

2.6 Summary

The 1980 NMAB report expressed the "need for continued long-range research in material deformation at high strain rates, in order to extend predictive capabilities. Such research should emphasize an understanding of micromechanical mechanisms of dynamic deformation within a thermodynamically and kinematically consistent large-deformation theory, based on macroscopic dynamic experiments and observation of resultant microstructure." While indeed much progress has been made by the materials science and mechanics communities in the past decade in terms of theories for dislocation motion and arrangement under dynamic straining, microvoid nucleation and growth, the roles of heterogeneities and thermal softening in shear banding, and fragmentation/communition, most of this work has not yet been effectively transferred into solution methodologies for impact problems. Furthermore, some important material behaviors such as anisotropic deformation and damage have not really been addressed. Strong links between micromechanical material modelers who work closely with material testing and characterization researchers and hydrocode developers appear to be lacking, except in isolated cases. The importance of the link of material model development with hydrocodes will be discussed in the next section.

In summary, key aspects of dynamic material behavior which must be addressed by a material model include:

- (i) temperature- and rate-dependence of plastic flow and of material hardening,
- (ii) dynamic recovery effects,
- (iii) effects of textural and dislocation substructure-induced anisotropy,
- (iv) solid state and solid-liquid phase transformations,
- (v) appropriate additional, irreversible deformation mechanisms such as twinning,
- (vi) damage evolution,
- (vii) failure criteria,
- (viii) anisotropic effects of damage, shear banding and transformations,
- (ix) coupling of deformation and evolution of various salient forms of material damage, and
- (x) a thermodynamically and kinematically consistent framework for pressure-dependent, finite deformation elasto-viscoplasticity.

At present, only certain of these elements appear in models commonly employed in hydrocodes for dynamic impact analysis. In particular, items (i), (vii) and the kinematical

aspect of (x) are common to numerous models. Items (vi) and (viii) are considered by certain models such as the SRI NAG-FRAG algorithms, but are not as common. Item (ix) is rarely considered in models for metallic behavior, but is typically considered in recent models which describe the behavior of ceramics (cf. Krajcinovic & Fonseka, 1981; Margolin, 1983; Costin, 1983; Chen, 1991). Items (ii), (iii), (iv), (v), (ix) and the thermodynamic coupling aspects of (x) have been essentially ignored in penetrator impact analyses with a few exceptions.

Item (x) is important for two reasons. First, the evolution of internal state must result in dissipation of energy consistent with experimentally observed values inferred from temperature rise during, for example, adiabatic deformation. Even if some phenomenological form for dissipation is assumed in the energy equation (e.g., 90-95% of work of inelastic deformation dissipated), the level of dissipation revealed by a thermodynamically consistent formulation yields insight into the physical consistency of the model. Second, much is now known regarding the geometry of crystallographic slip and its polycrystalline generalization (cf. Rice, 1971; Asaro, 1983; Asaro & Lowe, 1985; Weng, 1987; Bassani, 1990; Cuitino & Ortiz, 1992). Macroscopic continuum theories must reflect this knowledge appropriately through adoption of substructure spin concepts based on crystal plasticity results (cf. Dafalias, 1983, 1985, 1987; Aifantis, 1985, 1987; Bamann, 1990; Cleja-Tigoiu & Soos, 1990; Aravas & Aifantis, 1991; Miller & McDowell, 1992). The usual kinematics of finite deformation as embodied in the algorithm of Wilkins (1963) must be modified accordingly when sources of material anisotropy due either to deformation or damage are present.

In many cases, the technology for reflecting microscale effects such as phase transformation, anisotropy, etc. has not yet been developed to a level which facilitates analysis using macroscopic continuum models. It is clear, however, that such inclusive models require a more precise definition of material state than can be offered by simple models which are presently employed, i.e., equations (5)-(10). As outlined in the next section, the need to analyze impact problems using computationally efficient hydrocode algorithms constrains the complexity of these constitutive relations. As will be discussed in Section 4, internal state variable models offer potential to address these ten model requirements within a common framework (cf. Miller, 1987), yet preserving a degree of simplicity appropriate for solving impact problems.

3. HYDROCODES AND SIMULATION OF SHOCK WAVE DAMAGE

Fundamentally based on the decomposition of the dilatational and deviatoric relations introduced by Wilkins (1963), finite difference and finite element codes have been developed to simulate a wide range of impact and blast conditions of interest. As an historical artifact of the approximate hydrodynamic analyses of penetrator impacts in the 1940s and 1950s, these codes are termed "hydrocodes" in spite of their applicability far beyond hydrodynamic conditions. Generically, this term has come to encompass general purpose explicit integration schemes, both finite difference and finite element, used in dynamical problems of deformable media with shock waves.

Large deformations are a hallmark of ballistic impact problems. Several generations of codes, applying either the Lagrangian or Eulerian descriptions of deformation, have been developed since the 1960s. The state-of-the-art in constitutive equations for deformation and damage under dynamic loading conditions has simultaneously advanced; these improvements have been incorporated, to a considerable extent, in the hydrocodes (particularly in Lagrangian codes). The reader is referred to a review by Johnson and Anderson (1987) for details of the historical evolution of hydrocodes.

It is fundamentally essential to include a discussion of the hydrocodes in any thorough commentary on the state of modeling of impact and blast wave damage. In particular, the implementation within the numerical scheme alters, in most cases, the effect of the constitutive equations; although the constitutive equations are strictly distinct from the equations of motion, conservation of mass, and conservation of energy, solution algorithms act to modify the role of constitutive equations to some extent. In most cases, this coupling owes to the fact that the minimum mesh or grid dimension in hydrocode applications is large compared to the gradients associated with interfaces or shock waves. Otherwise, three-dimensional shock calculations would be unrealistic, even for the most powerful serial or parallel computers.

3.1 Lagrangian and Eulerian Hydrocodes

The choice of Lagrangian or Eulerian coordinates has a profound impact on the solution methodology. Lagrangian coordinates which are fixed in the material are commonly employed for deformable solids, with either the initial, undeformed state as the reference configuration (e.g., elastomers) or the current state updated at each time step as a reference for the next increment of deformation (i.e., updated Lagrangian). In this case,

nodal points are attached to material points; hence, the finite difference or finite element mesh deforms along with the material. This approach has the advantage of tracking material deformation history and orientation of material elements in a very straightforward manner. However, during large deformations the mesh distortion can eventually compromise solution efficiency and accuracy, necessitating a rezoning of the mesh to achieve more reasonable aspect ratios of elements. This rezoning procedure is nontrivial since it must obey conservation of mass, momentum, total energy and the constitutive relations.

Eulerian coordinates are commonly employed in fluid mechanics and assume a configuration or mesh of nodal or grid points which are fixed in space. The material flows through the mesh, and convective terms must be included in all rate quantities to account for mass transport. There is no need to rezone since the mesh does not distort with deformation, but other problems arise associated with tracking of material memory and interfaces.

In addition to the deviatoric and hydrodynamic constitutive laws for the materials involved, the set of equations to be solved under adiabatic impact conditions is as follows:

Eulerian: Conservation of

$$\text{mass: } \frac{\partial \rho}{\partial t} + v_i \frac{\partial \rho}{\partial x_i} + \rho \frac{\partial v_i}{\partial x_i} = 0$$

$$\text{linear momentum: } \frac{\partial v_i}{\partial t} + v_j \frac{\partial v_i}{\partial x_j} = f_i + \frac{1}{\rho} \frac{\partial \sigma_{ij}}{\partial x_j} \quad (11)$$

$$\text{energy: } \frac{\partial e}{\partial t} + v_i \frac{\partial e}{\partial x_i} = f_i v_i + \frac{1}{\rho} \frac{\partial (\sigma_{ij} v_i)}{\partial x_j}$$

where ρ is the material density, v is the velocity, e is the specific total energy (i.e., specific kinetic plus internal energies, $e = 1/2 v_j v_j + E$), σ is the Cauchy stress tensor, f is the external body force vector (e.g., gravity), and x are the spatial coordinates.

Updated Lagrangian: Conservation of

$$\text{mass: } \frac{D\rho}{Dt} + \rho \frac{\partial v_i}{\partial x_i} = 0$$

$$\text{linear momentum: } \frac{Dv_i}{Dt} = f_i + \frac{1}{\rho} \frac{\partial \sigma_{ij}}{\partial x_j} \quad (12)$$

$$\text{energy: } \frac{De}{Dt} = f_i v_i + \frac{1}{\rho} \frac{\partial (\sigma_{ij} v_i)}{\partial x_j}$$

Clearly, the distinction between the Eulerian and updated Lagrangian approaches lies in the use of convective terms in the former; in the latter, the coordinates are allowed to deform with the material, and the current state is set as the reference for the next increment of deformation. Since the increment of deformation at each time step is infinitesimal, the Cauchy stress may be regarded as a suitable approximation of the second Piola-Kirchhoff stress tensor with respect to the updated reference configuration.

In the energy equation, electromagnetic and chemical effects have been ignored. Moreover, it is typically assumed that the process is adiabatic such that heat conduction may be excluded from the energy equation. Accordingly, the temperature in the Lagrangian description is most often assumed to depend linearly on the specific internal energy per unit reference volume, E , i.e.,

$$T = T_0 + \frac{E}{C_v \rho_0} \quad (13)$$

where ρ_0 is the density in the reference state and C_v is the specific heat.

In view of the contribution of dynamical effects and the short time duration of impact events, explicit integration is commonly employed, with the size of the time step for a Lagrangian analysis governed by the Courant stability criterion (Zukas et al., 1981)

$$\Delta t \leq \frac{kL}{c} \quad (14)$$

where L is the minimum mesh dimension in the computational grid and c is the speed of sound; scalar k typically lies in the range $0.6 < k < 0.8$. As discussed by Anderson (1987), this time step must be reduced further if artificial viscosity is introduced to dampen oscillations resulting from the passage of compressive shock waves as discussed later.

Lagrangian hydrocodes (e.g., EPIC and DYNA) employ user-designated slidelines to represent interfaces to enforce continuity of pressure, boundary tractions and/or normal velocity, but such analyses must be periodically rezoned or remeshed for very large deformations to avoid excessive numerical error. Moreover, the time steps become very small as the mesh progressively distorts. For ratios of plate thickness to penetrator diameter in excess of ten, for example, 30-50 rezone may be necessary in Lagrangian analyses; as a result, a "semi-Eulerian" flavor may emerge, with diffusion of material history effects due to the approximate nature of the rezoning scheme. Another scheme to inhibit the excessive mesh distortion in the vicinity of a long rod penetrator is to admit erosion of the contact surfaces, which is an observed feature of penetration (cf. Kimsey & Zukas, 1986; Johnson & Stryk, 1987). An erosion criterion is specified by the user and, when met, element strength attributes are degraded while the mass is retained at eroded nodes, permitting subsequent dynamical interaction with the intact bodies. The erosion criterion is typically based on a critical level of effective plastic strain in an element (Johnson and Stryk, 1987). The physical basis of such a criterion is questionable since no specific failure mechanisms are identified other than some form of comminution.

An advantage of the treatment of material interfaces as slidelines in Lagrangian codes is the capability to specify frictional or erosion conditions; treatment of interfaces offered by discrete slidelines has application in fluid-structure interactions, contact between bodies involving friction and penetration of targets. A significant advantage of Lagrangian codes is their capability to track damage and other internal state variables very accurately with respect to deforming material. Accordingly, numerical diffusion problems associated with material convection inherent in Eulerian approaches are avoided.

An alternative method under development is the coupling of Eulerian and Lagrangian approaches, which may be accomplished in several ways. One way is to employ an Eulerian mesh for part of the domain (e.g., fluid) and a Lagrangian mesh for the remainder (e.g., solid); in this way, the key advantages of each methodology can be exploited. Of course, the interface or link between these two different meshes requires some assumptions and has not yet been fully automated (cf. Dike et al., 1993) for arbitrary interface geometries. Another way is to use both Eulerian and Lagrangian meshes,

admitting interface cells or gradient conditions at the interface; this is the basis for ALE (arbitrary Lagrangian-Eulerian) hydrocodes presently under development. Such hydrocodes may offer the most robust solution for a wide variety of impact and blast wave phenomena.

The Eulerian approach is very convenient since it requires no rezoning or user intervention during the solution. Existing codes are predominately based on finite difference algorithms (cf. McGlaun & Thompson, 1990). The approach is well-suited to hypervelocity penetration where the extreme deformation is confined to the vicinity of the penetrator and is particularly useful for penetration of thick targets ($t/D > 3$). However, it leads to diffusion and smearing of interfaces as well as other variables such as effective plastic strain or damage due to the fact that the property gradients near interfaces as well as gradients of material damage/deformation are steep within cells in many cases, and material convection is very difficult to account for with a high degree of accuracy in a relatively coarse mesh. Accordingly, Eulerian analyses are somewhat less effective for thin targets which deform significantly over a large area. Eulerian analyses make use of so-called "mixed cells" at boundaries to represent the presence of a boundary or interface between two different substances. This approach can be inaccurate if the densities between the two substances differ greatly, for example gas-solid interfaces (cf. Smith, 1983).

The treatment of the convective terms in the evolution of stress and internal state variables is nontrivial since only a fraction of material within a cell convects to the next cell during a small time increment typical of explicit codes (the so-called advection problem). Tensorial variables appear to offer particular challenge in this regard. Typically, material convection effects are handled by permitting a Lagrangian distortion over the time step, followed by a rezone to the original grid configuration, including appropriate material transport across cell boundaries (Zukas et al., 1981; McGlaun & Thompson, 1990). To inhibit diffusion of material, massless tracer particles are introduced which trace cumulative paths across which material transport is not permitted. More global rezoning may also be employed in order to maintain a fine mesh in the region of greatest interest, although this is somewhat uncommon. Second order advection algorithms along with a rule of mixtures approach for mixtures of multiple phases within cells weighted according to the volume or mass fraction of phases (cf. McGlaun, 1992) are routinely employed in the modern generation of Eulerian hydrocodes such as CTH or MESA. The plastic strain increment is assumed to be the same in all phases within a cell in order to ensure consistency in the definition of the total strain, yet another approximation. As a result of these algorithms,

numerical smearing or diffusion may occur, even resulting in anomalous results such as reduction of damage ("numerical healing") which are not physically plausible.

In Eulerian hydrocodes, failed material is typically treated as a void within a cell, thereby removing its mass from subsequent interaction with the contacting bodies, another approximation. The problem of material convection has apparently limited the implementation of sophisticated evolutionary measures of internal material state in Eulerian wavecodes. The averaging process of internal state within each cell poses a problem for multi-phase cells or cells which are only partially damaged (cf. Predebon et al., 1991). The solution process consists of three phases, i.e., (i) solve the difference equations for momentum and energy without convective terms to get velocity and internal energy, (ii) convect mass according to velocity and density of the cell and of neighboring cells, with the moving mass carrying its share of the cell momentum, energy, stress, etc., and (iii) update the velocity and internal energy accordingly. Sliding and eroding interfaces can also be introduced, albeit perhaps less directly, in Eulerian codes; for hypersonic, low velocity impacts, Silling (1992) has developed a boundary layer approach which smears the effect of the interface in the CTH simulation over several cells along the interface, resulting in more realistic analyses.

3.2 Artificial Viscosity

Another important feature of essentially all hydrocodes is the introduction of an artificial bulk viscosity to smear the shock wave over several elements, thereby inhibiting numerical instabilities associated with excessively steep stress and velocity gradients (cf. Wilkins, 1976). The thermodynamic pressure P_0 is obtained by adding a strain rate-dependent component, q , to the hydrostatic pressure, i.e.,

$$P_0 = P(\rho, E) + q$$

$$q = c_1 \rho L c |\dot{\epsilon}| - c_0 \rho L^2 \dot{\epsilon} |\dot{\epsilon}| \quad \text{for } \dot{\epsilon} < 0, \quad q = 0 \quad \text{for } \dot{\epsilon} \geq 0 \quad (15)$$

$$\dot{\epsilon} = \frac{\dot{V}}{V} = \text{rate of dilatation}$$

where L is a characteristic length of the grid, c_0 and c_1 are constants, and c is the local speed of sound (cf. Zukas et al., 1981). It is not uncommon for the peak pressures to be

underestimated significantly in analyses with artificial viscosity as a result of the smearing. Using deviatoric rate-independent plasticity, artificial viscosity changes the governing differential equations to parabolic from hyperbolic, so that localization does not follow the characteristics of the hyperbolic wave equation. Also, the time step must decrease when artificial viscosity is added in order to meet the Courant stability criterion. Artificial viscosity is essentially a uniform feature of Lagrangian hydrocodes, but Eulerian codes may not require it if damping is implicit in the discretization scheme (Zukas et al., 1981), although the latter is also an artificial source.

Hourglass or keystone viscosity is often additionally introduced to suppress hourglass distortion instabilities behind the shock wave which commonly occur with quadrilateral elements; hourglass viscosity alters the solution in a somewhat physically inconsistent manner and should be used only when necessary to inhibit such non-unique modes of deformation (cf. Belytschko et al., 1984).

Alternatively, shock viscosity formulations which more faithfully represent material viscosity effects may be preferable to artificial viscosity (Anderson, 1987), particularly if the details of the shock stress and strain fields are essential for micromechanical damage formulations. This may be accomplished without bulk viscosity (Swegle & Grady, 1985) by introducing a rate-dependent deviatoric flow rule, e.g.,

$$\dot{\tau} = 2G(\dot{\epsilon} - A'\tau_v^2) \quad (16)$$

where τ_v is a viscous stress and A' is a constant. This linear hypoelastic expression is analogous to the law

$$\dot{\tau} = 2G(\dot{\epsilon} - \dot{\epsilon}^v) \quad (17)$$

where $\dot{\epsilon}^v$ is the viscous strain rate. Even though the shock is smeared by the actual viscosity in this case, it is much more localized than in the case of artificial viscosity and highly refined meshes (e.g., 10^{-3} to 10^{-2} mm) may be required to resolve the shock. From a theoretical viewpoint, much of the need for artificial viscosity arises from the common use of a rate-independent material law (e.g., perfect plasticity) for the deviatoric plasticity, in contrast to the strong rate-dependence which is typically observed in metals. From a practical viewpoint, however, artificial viscosity may be required to supplement deviatoric viscosity to further stabilize the shock wave solution for a coarse mesh. However, the

implications of artificial viscosity on the post-localization behavior of rate-dependent materials behavior are not fully understood.

3.3 Modeling of Material Damage

Considering a range of problems ranging from low velocity impacts (< 500 m/s) to ordnance velocity and above, it is clear that material deformation and deformation rates vary accordingly. At lower velocities, relatively small strains and strain rates result, suggesting the applicability of Lagrangian codes in view of their superior treatment of interfaces and history effects. Eulerian codes may have advantages for modeling long rod penetration at ordnance velocities and above; such codes are useful for phase transformations, impact cratering in long rod penetration and estimation of debris clouds during hypervelocity impact (Johnson & Anderson, 1987).

For impact velocities between about 1.5 and 4 km/s, material deformation and damage effects, including history dependence, play a potentially very strong role. Between 4 and 8 km/s, however, material strength plays a progressively weaker role as density becomes paramount in the hydrodynamic regime of interpenetrating fluids. Even under hypervelocity conditions, material strength effects play a strong role in cratering behavior. Moreover, accurate prediction of penetration hole size for long rods and shaped-charge jets at hypervelocity requires consideration of material damage evolution (cf. Raftenberg, 1992a). At low impact velocities (e.g., near the ballistic limit), shear plugging or extensive plastic deformation may be key failure modes. For $V > 4$ km/s, material separation may occur via several competing mechanisms, e.g., spallation, plugging, radial cracking, ductile hole growth and fracture, extensive flow, erosion, shear banding, melting, etc. The treatment of material damage in most hydrocodes is presently at a relatively rudimentary level. Most codes rely on instantaneous rather than evolutionary damage/failure criteria. For example, a spall stress is specified for spallation. An accumulated effective plastic strain criterion, at most additionally dependent on the hydrostatic stress and temperature, is specified to designate failure by void growth/coalescence or other ductile fracture mode (Johnson & Cook, 1983, 1985). A critical effective plastic strain and temperature combination has been specified for shear banding (cf. Raftenberg, 1992a). Other common criteria include volumetric strain with a tensile cut-off pressure. Often, partial failure is admitted, such that the damaged solid behaves as an inviscid fluid incapable of supporting shear or tensile stresses. Such instantaneous or semi-evolutionary criteria, loosely linked to physical processes, are attractive in view of the short time-scales associated with the impact problem. However, it must be emphasized that strong path history effects may

occur in impact problems since damage can evolve at the rate of cumulative inelastic strain. Moreover, effects of multiple stress wave reflections introduce an inherent history dependence to damage evolution and its coupling with the stress-strain response of the material. Even in hypersonic impacts, reflected stress waves can result in spallation and fragmentation (cf. Raftenberg, 1992b) following the primary penetration event.

Limited studies have verified the importance of evolutionary measures of damage in modeling actual impact processes. In particular, the coupling of progressive damage with the stress-strain behavior of the material appears to be of first order importance. As pointed out by Seaman (1975), there is no instantaneous jump from undamaged to fully separated material; damage grows as a function of applied stress and time, the wave propagation is affected by the stiffness reduction due to damage, and even small, incipient damage levels are pertinent since they can facilitate damage evolution and localization. It is also pointed out that instantaneous criteria may be appropriate at very high rates, but this postulate has not been fully verified. Certainly, impact phenomena in the vicinity of the ballistic limit and below will generally involve significant contribution from evolution of damage as opposed to instantaneous damage. Furthermore, the practical problem of perforation of plates, in contrast to that of the semi-infinite target, is expected to depend more significantly on the evolutionary representation of damage in view of the finite dimensions and associated wave reflections, large scale structural oscillations, etc.

As reviewed later in this report, Bammann et al. (1993) at Sandia National Laboratories in Livermore have demonstrated through experimental studies and DYNA simulations of a cylindrical aluminum projectile impacting an aluminum plate in the vicinity of the ballistic limit that use of an evolutionary micromechanically based void growth law was necessary to model the level of shear localization (incipient plugging) and cracking behavior which was observed. A damage measure based on cumulative inelastic strain differed significantly from the void growth model in predicting shear localization (incipient plugging) and damage using DYNA. In contrast to the vast majority of studies reported in the literature, the parameters in the void growth model (e.g., initial void density) were determined by quasi-static experiments on circumferentially notched tensile specimens which were completely independent of the impact problem. In addition, the internal state variable model for viscoplastic behavior developed at Sandia was employed (cf. Bammann, 1990), which is known to offer more realistic treatment of path history dependence of finite inelastic deformation than, for example, the Johnson-Cook model which is commonly employed in codes.

Very little use of such deformation-coupled evolutionary damage laws has been reported in the literature for long rod penetrators, shaped-charge jets, etc. Hertel (1992) reports the performance of the Eulerian hydrocode CTH in modeling normal and oblique long rod penetration and oblique impact of a copper sphere on a steel plate. However, it is not clear that the choice of material parameters for damage are independent of the analyses in these cases. Similar analyses have been conducted for other Eulerian hydrocodes (cf. Mandell & Henninger, 1992; Mandell, 1993a, 1993b). Some comparisons have been made, for the same material laws, between Eulerian and Lagrangian hydrocodes. For example, Predebon et al. (1991) compared Lagrangian (HEMP) calculations with Eulerian (CSQ) for the Taylor problem of cylinder impact on a rigid half space with and without work-hardening. The Eulerian code, for comparable solution times, significantly underpredicted the inelastic strain in the vicinity of the leading edge of contact due to averaging of the plastic strain over the cell. A finer mesh in the Eulerian code gave comparable results, but at much higher computational cost. It therefore seems clear, in view of problems of treating interfaces and evolutionary damage measures in Eulerian codes, that it is prudent to properly define classes of impact problems which may be treated with such codes, with long rod penetration being one possibility. More than likely, ALE codes or links between Eulerian and Lagrangian codes will best exploit the outstanding features of each.

As stated by Zukas et al. (1981), "The principal limitation of computer codes for high velocity impact studies is the uncertainty in the material response as regards failure...Despite the fact that material failure is a time-dependent process, most production calculations are performed with simple, empirical, time-independent criteria based on maxima and minima of field variables (i.e., maximum tensile stress, maximum shear strain, maximum plastic work, relative volume)." These assertions basically echo the findings of the NMAB Committee report (1980). Although some progress has been made in the last decade, it is clear that much research on microscale failure mechanisms under dynamic loading has not yet translated into macroscale models which are useful, practical and sufficiently accurate for production codes.

As suggested by Zukas et al. (1981), there are several possible reasons for this lag of implementation which have apparently persisted. First, in the face of the apparent complexity of the various failure modes and their interaction, use of sophisticated equations for deformation and damage may be unwarranted if they are not fully verified. Second, the cost of material characterization for more sophisticated laws is prohibitive. The latter

problem is becoming perhaps less troublesome in view of the plethora of testing facilities and quantitative analysis facilities which exist both at national laboratories and at universities. The first qualification has merit in view of the fact that the numerical approximations made in hydrocodes actually modify, to some extent, the role of the constitutive equations. Hence, continued development of hydrocode algorithms with emphasis on treatment of material history effects and failure processes must accompany any serious effort to implement more realistic constitutive laws. This coupling of material model development with that of hydrocodes was not fully appreciated or expressed in the 1980 NMAB report, for example. In fact, the issue of hydrocode algorithms is indeed difficult to separate from development of material models for interactive, nonlinear phenomena.

A final comment pertains to the lack of introduction of material anisotropy in hydrocodes for impact analyses. Anisotropy in penetrators and RHA materials forms predominately in response to deformation since most materials of interest are initially isotropic. As outlined in an earlier section, large strains can produce significant textural and microstructural anisotropy, leading to as much as a factor of two error in cumulative strain using an isotropic hardening plasticity theory, depending on the state of stress. Moreover, the process of shear localization/banding results in local anisotropy. Phase transformations are very similar in terms of their effect, producing a directional attenuation to stress waves and different deformation-damage coupling. Furthermore, some penetrators may already be textured due to processing, resulting in initial anisotropy. Very little is known of the precise nature of the effects of anisotropy on penetration since such analyses are rarely considered in the literature. However, there is a definite prospect of altering the effectiveness of penetration through introduction of material anisotropy which remains largely uninvestigated. In view of the significance of anisotropic effects in quasi-static experiments, it appears fruitful to pursue their implementation in hydrocodes to assess the potential effect on impact and blast wave damage.

3.4 Equation of State for Hydrostatic Response

The 1980 NMAB report did conclude, however, that the current equation of state formulations were largely satisfactory, apart from consideration of phase transformations. However, the solution may be significantly dependent on the equation of state selected. Many hydrocodes make use of a pressure-dilatation-internal energy equation of state which is held to be valid for shock loading conditions, e.g.,

$$P = \hat{P}(E, v) = -\frac{\partial \Psi(v, T)}{\partial v} \quad (18)$$

where v is the specific volume ($= 1/\rho$), P is the pressure, E is the internal energy per unit of undeformed material volume, and Ψ is the Helmholtz free energy. Under the assumption that the Gruneisen parameter

$$\Gamma = v \left(\frac{\partial P}{\partial E} \right)_v \quad (19)$$

is independent of pressure, the pressure can be evaluated, given Γ , for any other pressure off the experimentally measured Hugoniot curve, i.e.,

$$P = P_h(v) + \frac{\Gamma}{v}(E - E_h) \quad (20)$$

where the "h" subscripts refer to the Hugoniot reference curve. A common form used in application of this concept in hydrocodes is given by

$$P = (K_1\mu + K_2\mu^2 + K_3\mu^3) \left[1 - \frac{1}{2}\Gamma\mu \right] + \Gamma(1 + \mu)E \quad (21)$$

where K_1, K_2, K_3 , and Γ are constants,

$$\mu = \frac{\rho}{\rho_0} - 1 \quad (22)$$

is the compressibility of the material, and ρ_0 is the density in the reference state. More generally, a polynomial form in μ and E may be fit for a given material (cf. Hallquist, 1988).

For metals, the elastic deviatoric strains are typically small while the elastic dilatation can be rather significant. Accordingly, the decomposition into a small strain, linearized deviatoric elastic relation along with a finite strain (hyperelastic) formulation for

the pressure volume relationship has persisted. As shown by Scheidler (1993), however, this is generally inconsistent. The deviatoric stress must be dependent on the pressure in order to achieve a thermodynamically consistent elasticity formulation for either isothermal or adiabatic processes. As a consequence, shear stresses in the vicinity of the HEL can be significantly altered relative to those based on pressure-independent elastic constants. The equation of state formulations which are prominently used in hydrocodes are based on behavior of fluids rather than that of solids. The Gruneisen equation of state holds reasonably well for single phase liquids and metals, but is not applicable to porous materials (compressive dilatational effects), vaporization, phase change, or combustion processes (cf. Asay & Kerley, 1986). Hence, in addition to the lack of consideration of phase transformations or of damage in most of these formulations, a consistent formulation for hyperelastic behavior of solids is lacking. It should be kept in mind, however, that the accuracy of the equation of state may often be of secondary importance in the low velocity impact regime.

3.5 Summary: Hydrocodes as an Engineering Tool for Impact Evaluation

In view of the complexities of material deformation mechanisms and the implementation of approximate material laws in hydrocodes, robust and accurate predictive capability has not yet been fully demonstrated for long rod penetration, shaped-charge jets, and impacts involving frictional interfaces. On the other hand, good correlations with ballistic impact experiments have been achieved by "tuning" a small number of parameters related to a dynamic yield strength and scalar failure criteria such as cumulative plastic strain, material erosion criterion, etc. In fact, failure criteria are commonly used as adjustable parameters to fit experimental results. Each new penetrator/armor system requires such adjustments to correlate with experiments when simple deformation and failure laws are used. Such an engineering approach was supported by the NMAB report (1980) which recommended to "employ an iterative procedure of successive refinements using computations with simple failure and flow laws along with ballistic testing to produce useful results for many applications. Refine material properties/characterization and computational/material models accordingly and repeat."

There are certain advantages and disadvantages of the emphasis on iterative correlation and associated parameter fitting using very simple, isotropic theories. The chief advantage is that once a penetrator-target system is characterized over a range of velocities, a wide range of impact conditions (e.g., velocity, obliquity, geometry) can, in principle, be evaluated with some confidence, presumably for purposes of lethality analysis. Moreover,

such analyses can be conducted rather efficiently since they are based on very simple idealizations. Even a novice analyst with limited background in mechanics/materials can rapidly assimilate the theory and parameter sets involved.

However, there are disadvantages which may prove to be equally compelling in the post-cold war era. The philosophy of using simple failure criteria as the basis of correlating parameters has underpinned ballistic search and analysis programs since the 1950's, and has continued with the introduction of high speed computers. Although it is a practical approach, it depends on significant levels of ballistic testing, particularly full scale testing. Over the past few decades, emphasis has been placed on testing scaled down penetrators since delivery requirements are reduced. The applicability of small scale penetration tests to examine full scale effects was addressed by Anderson et al. (1992) by forming nondimensional parameter sets of all pertinent variables and using the Buckingham-Pi Theorem for similitude. It was shown that material responses such as strain rate effects on flow stress and damage evolution which are nonlinear in time or strain rate lead to lack of similitude. However, the lack of similitude diminishes as the projectile overwhelms the target. Analyses were performed for a tungsten-alloy penetrating 4340 steel hypersonically at 1.5 km/sec with L/D = 12 for both a prototype and model where the model was scaled down from the prototype by a factor of 10. The CTH code was used, and it was assumed that all of the plastic work goes into thermal heating. The Johnson-Cook model was used for the target material while a rate-independent plasticity law was assumed for the penetrator. The analysis did not exhibit similitude expected on the basis of the Pi Theorem. The material strain rate dependence induced a 5% disparity between depth of penetration for the two cases; this is small compared to scatter in experimental data and may explain the observed independence of the depth of penetration on the size scale. The Eulerian code CTH provides for "mixed cells" at the interface of the penetrator and target materials which provides for erosion. No failure criterion was included for the inelastic deformation of the material so the behavior is unrealistically compliant and no fragmentation occurs. Detailed analyses were provided for penetrator erosion and velocity of both the leading edge and back edge throughout the impact. Although not included in the simulations, effects of void nucleation and growth are another possible source of discrepancy in scaling.

The lack of similitude in scaling from "laboratory" to full scale penetrators is likely due to the fact that material rate-dependence, temperature-dependence and damage behavior must be described by a more extensive, physically detailed set of variables (cf. Jones,

1989). It is particularly troublesome in the vicinity of the ballistic limit. With ever-decreasing emphasis on ballistic testing (particularly full-scale) at ordnance velocities, it seems clear that increased emphasis must be placed on simulation. Accordingly, identification of material mechanisms and their representation in appropriate constitutive laws must be the focal point of future developments/improvements. Likewise, development of predictive capability must become increasingly important rather than a purely correlative posture. At the very least, such improvements in constitutive modeling will shed light on scaling discrepancies so that ballistic laboratory data can be properly interpreted.

There are two primary types of applications of models for material deformation and damage in hydrocodes:

- (i) Assessment of vulnerability/lethality of existing, experimentally characterized penetrator-target systems. This sort of application may rely heavily on the NMAB strategy of iterative testing, simulation, adjustment, etc. and may be required to deliver results of high confidence. This approach has dominated the traditional design and development cycle of penetrator-target systems. However, it may be difficult to sustain this strategy in the future due to limited resources.
- (ii) Design/development of new penetrator-target systems. This type of application, heretofore primarily in the domain of processing science and ballistic testing, will most likely demand increasing emphasis on simulation prior to prototype development and testing. Advanced reactive and composite armor development has involved simulation at the beginning of programs to assess concept viability. Future technology breakthroughs will most likely be achieved with development of accurate micromechanical material models which are employed in the context of hydrocode simulations.

Given the myriad of possible material deformation and failure mechanisms, it is clear that the goal of a truly predictive methodology is a very demanding one. However, it is consistent with the NMAB report (1980) suggestion that research be accelerated in "micromechanical failure mechanisms, particularly in the high strain rate regime, in order to provide the basis for development of adequate failure models." The possibility of accurate simulations for purposes of design and development of new penetrator/target concepts and systems depends heavily on development of micromechanical models for deformation and failure modes embedded within a thermodynamically and kinematically consistent large-deformation theory. To date, theory for hydrocodes has followed the framework set forth by Wilkins (1963), as mentioned at the outset of this section. Attention has been devoted

to the equation of state only as it regards the relation between hydrostatic pressure and dilatation, with little consideration for consistency at low pressure. Moreover, plasticity has been treated predominately in terms of observable variables only. In a minority of cases, progressive damage is coupled with evolving deformation rather than just assigning loss of stiffness at failure. Due to these limitations, particularly the absence of explicit expressions for damage in the evolution equations for the dynamic elasto-viscoplastic response, the goal of developing adequate failure models based on inclusion of micromechanical failure mechanisms has been difficult to achieve. In the next section, we will discuss a theoretical framework which offers a more complete characterization of the state of the damaged material, consistent with much of the research within the last 15 years on internal state variable theories for finite deformation.

4. AN INTERNAL STATE VARIABLE FRAMEWORK FOR MATERIAL BEHAVIOR

It is clear from the preceding discussion that there are several very important features of dynamic material behavior during impact which must be addressed by a comprehensive continuum framework which couples modes of deformation and damage. The implementation of various mechanisms of damage into constitutive equations has conventionally been achieved in an ad hoc fashion, dependent on penetrator/target combination and impact velocity. While this state of affairs may be reflective of the present level of understanding of the contribution of different failure mechanisms, it is unsatisfactory from a fundamental standpoint in terms of developing simulation capability which can address (i) the shortcoming of scaling laws in extrapolating small scale impact experimental results to full-scale penetrator/target scenarios, (ii) the coupling of different damage modes, and (iii) the need for a truly predictive capability.

To address these issues, we consider an internal state variable framework at the macroscale (cf. Germain et al., 1983; Allen, 1991) which includes:

- (i) a viscoplastic flow rule with strain rate- and temperature-dependence, additionally depending on pressure, damage and initial anisotropy,
- (ii) damage-coupled elasticity,
- (iii) damage evolution laws for void and microcrack nucleation and growth, shear banding, phase transformations, etc., based on micromechanical solutions whenever possible,
- (iv) combined nonlinear isotropic-kinematic hardening with dynamic recovery effects, shock hardening effects (strain rate-dependence of hardening), etc.,
- (v) effects of textural anisotropy at large strains,
- (vi) effects of anisotropy associated with formation of deformation substructure,
- (vii) appropriate failure criteria for material separation, including effects of anisotropy and fragmentation,
- (viii) a statement of dissipation associated with these various mechanisms, and
- (ix) a free energy function which accounts for effects of both damage and composition on the recoverable energy of the material.

It may be stated almost categorically that the constitutive laws employed in current hydrocodes, with few exceptions, do not address most of these criteria. For Eulerian hydrocodes, difficulties associated with advection of damage quantities, particularly

anisotropic variables, may prove difficult to overcome. On the other hand, updated Lagrangian codes are well-suited to admit nearly all of the above features. Highly localized failure (softening) mechanisms such as shear banding require special consideration, much of which awaits clarification on the basis of micromechanical modeling.

Numerous internal state variable theories have been proposed and implemented in the last decade which include at least some of these attributes (cf. Miller, 1987). Perzyna (1984) has incorporated some of these elements (i, iii, vii and limited aspects of ii and vi) in a general framework for postcritical deformation and failure behavior of ductile metals. The model of Bammann (1985, 1990), implemented in DYNA, explicitly addresses at least seven of these requirements.

We next discuss the key elements of a local internal state variable theory for thermomechanical behavior of polycrystals.

4.1 Kinematical Decomposition

We adopt the multiplicative decomposition of the deformation gradient, \mathbf{F} , according to

$$\mathbf{F} = \mathbf{F}^e \cdot \mathbf{F}^n \quad (23)$$

where \mathbf{F}^e is the thermoelastic deformation gradient and \mathbf{F}^n is the deformation gradient associated with cumulative increments of plastic deformation and damage processes. This decomposition is a generalization of crystal plasticity concepts (cf. Asaro, 1983; Cuitino & Ortiz, 1992) to the polycrystalline case. Essentially, \mathbf{F}^n defines the rearrangement of the polycrystal associated with irreversible processes of dislocation motion and damage evolution (cf. Rice, 1971). Equation (23) is generally held to provide a more sound physical basis for elasticity of polycrystals at large strain than the assumption that elastic strains occur with respect to the initial, undeformed configuration of the body. In essence, the intermediate or so-called isoclinic configuration (cf. Mandel, 1971, 1973, 1974) defined by $\mathbf{F}^{e-1}\mathbf{F}$ serves as a reference frame for the thermoelastic deformation of the underlying atomic lattice; hence, the residual strain associated with the intermediate configuration has been achieved by introduction of lattice defects (point, line, microcracks, etc.) which persist after removal of applied stress. The precise interpretation of this intermediate configuration for a polycrystal is still open to further investigation as to

limitations (cf. Wilmanski, 1992), but it is rather well-accepted as a fundamental concept (cf. Dafalias, 1987; Cleja-Tigoiu & Soos, 1990). Elastic interactions of dislocations and defects at the microscale are reflected in F^n , while departure from this state due to temperature change and applied stress are reflected by F^e . The multiplicative decomposition of F may be further refined by introducing discrete inelastic effects to further subdivide F^n as discussed later. Moreover, linearization of thermoelasticity simplifies the expression for F^e . Along with the decomposition of free energy as discussed in 4.2, the multiplicative decomposition of F leads to an additive decomposition of the rate of deformation tensor in the current configuration, i.e.,

$$\mathbf{D} = \mathbf{D}^e + \mathbf{D}^n \quad (24)$$

4.2 Free Energy

The Helmholtz free energy (recoverable energy) is decomposed into the free energy associated with thermoelastic strain and temperature (modified by internal state variables), and microscopic elastic interactions due to the presence of defects according to

$$\Psi = \Psi^e(E^e, T, \xi_i, D_k, V_j) + \Psi^n(\xi_i, D_k, V_j; T) \quad (25)$$

where E^e is the thermoelastic Green strain with respect to the relaxed intermediate configuration $F^{e-1}F$ (cf. Bammann & Aifantis, 1987), T is absolute temperature, ξ_i is a set of internal state variables representing composition and the phases in the material, D_k is a set of damage internal state variables, and V_j are internal state variables associated with dissipative inelastic processes other than damage, such as dislocation arrangement. Variables ξ_i , D_k and V_j are expressed in the intermediate configuration since they evolve during F^n . For metals at low hydrostatic pressure, to first order, the intermediate configuration differs from the current configuration by at most a local rigid rotation since the thermoelastic stretch is typically small compared to unity under these conditions. However, at very high pressures common in impact problems, the elastic stretch can be significantly large, leading to elasto-plastic couplings in pushing forward the relations expressed in the intermediate configuration to the current configuration. Consequently, the evaluation of the thermoelastic strain must be conducted in the intermediate configuration. These internal state variables represent generalized displacement quantities. The elastic part

of the free energy function contains information necessary for developing the macroscopic elastic "equation of state," for both the deviatoric and hydrostatic stress-strain response.

According to standard arguments, this framework leads to certain relations for conjugate thermodynamic forces (cf. Germain et al., 1983; Lemaitre & Chaboche, 1985; Allen, 1991), i.e.,

$$\mathbf{S} = \bar{\rho} \frac{\partial \Psi}{\partial \mathbf{E}}$$

$$s = - \frac{\partial \Psi}{\partial T}$$

$$\mathbf{Y}_k = \bar{\rho} \frac{\partial \Psi}{\partial D_k} \quad (26)$$

$$\Sigma_i = \bar{\rho} \frac{\partial \Psi}{\partial \xi_i}$$

$$\mathbf{A}_j = \bar{\rho} \frac{\partial \Psi}{\partial V_j}$$

where \mathbf{S} is the second Piola-Kirchhoff stress with respect to the intermediate configuration, s is the specific entropy, and \mathbf{Y}_k , Σ_i and \mathbf{A}_j are generalized forces conjugate to the respective displacements. The barred density in equations (26) refers to the density in the intermediate configuration where these relations are written. Equation (26.1) for \mathbf{S} may be regarded as an equation of state which defines the nonlinear, temperature-dependent deviatoric and hydrostatic elastic relations.

The dependence of the free energy on composition permits treatment of certain types of isostructural phase transformations. Phase transformations which involve a change in crystal structure are more complicated. Each requires its own free energy function. This is important when addressing phase changes during penetration, for example.

4.3 Dissipation Potential/Internal State Variable Evolution

One standard approach used to specify the evolution of the inelastic strain and state variables in a consistent manner is to introduce a dissipation potential

$$\phi(S, Y_k, \Sigma_i, A_j; E^*, D_k, \xi_i, V_j, T) \quad (27)$$

The function ϕ is introduced from which the fluxes (rates of displacements) may be derived through the hypothesis of generalized normality, i.e.,

$$\begin{aligned} \bar{D}^* &= \frac{\partial \phi}{\partial S} \\ \dot{D}_k &= -\frac{\partial \phi}{\partial Y_k} \\ \dot{\xi}_i &= -\frac{\partial \phi}{\partial \Sigma_i} \\ \dot{V}_j &= -\frac{\partial \phi}{\partial A_j} \end{aligned} \quad (28)$$

Here, it is understood that the barred rate of deformation in (28.1) is the rate of change of the plastic part of the Green strain with respect to the intermediate configuration. The rate quantities in equations (28) are taken corotationally with respect to the intermediate configuration and satisfy material frame indifference (rotation invariance) requirements (cf. Dafalias, 1985).

The advantage of the assumption of generalized normality, although somewhat restrictive, is the unconditional satisfaction of the Kelvin inequality of the second law of thermodynamics, i.e.,

$$\text{intrinsic dissipation} = \bar{\phi} = S : \bar{D}^* - \sum (Y_k \dot{D}_k + \Sigma_i \dot{\xi}_i + A_j \dot{V}_j) \geq 0 \quad (29)$$

where summation is implied over all values of k , i and j , respectively, provided that ϕ is convex, always positive and contains the origin ($\phi(0) = 0$). Clearly, the term involving the

summation of force-flux products represents the rate of storage of nonrecoverable energy associated predominately with dislocation interaction stresses. This term is typically only 5% to 10% of the first term in the dissipation expression. The intrinsic dissipation enters directly into the adiabatic form of the local energy equation, i.e.,

$$\bar{\rho} \frac{D\mathbf{E}}{Dt} = \bar{\phi} \quad (30)$$

where certain second order thermomechanical couplings have been neglected (cf. Lemaître & Chaboche, 1985).

If a dissipation potential is not introduced, then constitutive equations must be introduced for each of the fluxes in equation (29). In this case, the dissipation inequality in equation (30) must be checked at each stage of the deformation process to ensure thermodynamic consistency. It should be noted that the common assumption that 90-95% of the plastic work is dissipated as heat implicitly takes into account the storage of energy due to damage and inelastic processes.

4.4 Failure Criteria

It is necessary to define the stage at each point in the material when failure is assumed to occur in terms of some physical mechanisms which relate to the state variables. Failure can be assumed to relate to either complete material separation (e.g., cracking) or partial loss of load-carrying capacity. For example, the material may lose resistance to tensile stress but may still support compression and/or shear. The failure condition may be expressed in terms of the generalized displacements, i.e.,

$$Z_d(\mathbf{D}_k, \xi_i, \mathbf{V}_j) = 0 \quad (31)$$

or generalized forces, i.e.,

$$Z_f(\mathbf{Y}_k, \Sigma_i, \mathbf{A}_j) = 0 \quad (32)$$

Often, very simple failure criteria can be selected as an approximation of complex interactions between failure modes which occur in practice. The evolution equations for the damage state variables account for much of the interaction since they are highly coupled.

4.5 Comments on the Overall Framework

This coupled thermomechanical framework has been used to treat a variety of problems in quasistatic elasto-plastic deformation and in evolution of damage. However, it has been infrequently applied to the simulation of ballistic impact.

The classical approach to elasto-plastic impact problems, based on the work of Wilkins (1963), does not introduce the free energy directly. The assumption of a linear elastic deviatoric response with an equation of state for the dilatational elastic response is very common, but is usually inconsistent with the thermodynamical framework discussed here. Since state variables are not commonly introduced for evolution of damage and internal state of the material, except for NAG-FRAG models and a few others used in specialized hydrocodes, most failure criteria are based on critical values of "external" variables such as cumulative inelastic strain or applied stress. The adoption of an internal variable approach implies a higher degree of understanding of the distinct mechanisms of material deformation and damage. To a large degree, much of this understanding has been only recently gained through application of micromechanics concepts and interaction of material scientists with applied mechanics researchers and computational specialists. Most of the development of these concepts lies ahead.

Either ductile or brittle failure processes can be treated within this framework. In fact, two major applications of micromechanical models to date have concerned void growth in ductile metals (cf. McClintock, 1968; Rice & Tracey, 1969; Oyane et al., 1973; Gurson, 1977; Needleman & Rice, 1978; Rousselier, 1981; Tvergaard, 1981, 1982; Doraivelu et al., 1984; Tvergaard & Needleman, 1984; Dragon & Chihab, 1985; Becker & Needleman, 1986; Needleman & Tvergaard, 1987; Kim & Carroll, 1987; Mear & Hutchinson, 1985; Lee, 1988; Cocks, 1989; Sun et al., 1989; Eftis & Nemes, 1991; Qiu & Weng, 1992; Narasimhan et al., 1992) and have focused on the form of plastic potential functions for pressure-dependent plasticity of porous solids. To varying degrees, microcrack nucleation & growth have been included in an internal state variable framework (cf. Seaman et al., 1985; Margolin, 1984; Chen, 1991; Mueller, 1992), although typically limited to small strain kinematics and linear elastic behavior. It must be stressed that the introduction of damage in either case contributes to both the F^e and F^n components of the total deformation gradient, in general, so that a consistent theory for large deformations must properly acknowledge elasto-plastic-damage couplings. If the damage-coupled elastic stretch is not infinitesimal due either to high pressure or effects of damage, then elasto-plastic couplings will naturally appear in the current configuration quantities pushed

forward from the intermediate configuration. Both Eulerian and updated Lagrangian codes are subject to this consideration.

Finally, the assumption of local action is invoked in the preceding internal variable framework. In other words, the effect of distributed deformation and damage mechanisms is assumed to be statistically homogeneous within the neighborhood of each point in question. However, when distribution or arrangement of defects and associated interaction of their effect influences the average behavior at the point, it may be necessary to explicitly take this distribution or arrangement into account. This may be accomplished either by effectively introducing length scales or gradients directly in the free energy and dissipation potential to modify the state variables, or by introducing gradient terms to modify stress-strain relations and the governing field equations. In either case, the theory becomes nonlocal in character, and size scales have a definite influence on the mechanisms of damage evolution and coalescence. Shear banding in low workhardening metals is a classic example of a nonlocal phenomenon in which the shear band thickness may be on the order of the computational mesh and is a key length scale in the problem. In contrast, typical dislocation-dislocation interactions in workhardening processes occur at length scales small enough to assure the viability of the assumption of statistical homogeneity. Fundamental advances are required in techniques which can produce averaged, and likely nonlocal, constitutive equations from micromechanical models of damage processes.

5. COMPUTATIONAL EXAMPLES USING AN INTERNAL STATE VARIABLE MODEL

In this section, we discuss some results of a specific internal variable model developed at Sandia National Laboratories in Livermore, CA which follows the format outlined in the previous section. This model is of interest because it has been implemented in the Lagrangian hydrocode DYNA and applied to a number of dynamic problems of relevance. Hence, it may serve as a simple example of the class of internal variable theories which reflect very specific deformation and damage mechanisms.

5.1 Sandia Internal State Variable Model

Sandia has developed and implemented a model for void growth in ductile metals within a state variable framework along the lines outlined in the previous section. This model has been used in blast wave and impact damage simulations (Bammann et al., 1993). Key features of the Sandia model include evolution of damage (porosity), coupling of the damage with both the elastic and viscoplastic responses, strain rate- and temperature-dependence, combined isotropic-kinematic hardening, and a kinematical decomposition based on an intermediate reference configuration for thermoelasticity.

The multiplicative decomposition $\mathbf{F} = \mathbf{F}^e \mathbf{F}^{th} \mathbf{F}^v \mathbf{F}^p$ is adopted, where \mathbf{F}^p is the deviatoric plastic deformation gradient associated with dislocation glide, \mathbf{F}^v is the dilatational deformation gradient associated with irreversible void growth, and \mathbf{F}^{th} is the deformation gradient associated with thermal expansion. A linearized form of the thermoelastic relations is assumed, so that \mathbf{F}^e does not explicitly involve thermal expansion, but it does depend on the degree of material damage in the form of voids. In the general case of nonlinear thermoelastic deformation, it would not be possible to split the thermoelastic deformation gradient. By standard arguments, we may write the rate of deformation in the following complementary manner

$$\dot{\mathbf{D}} = \dot{\mathbf{D}}^e + \dot{\mathbf{D}}^{th} + \dot{\mathbf{D}}^v + \dot{\mathbf{D}}^p \quad (33)$$

The damage-coupled, linear elastic stress-strain relation is given by

$$\mathbf{S} = \mathbf{C}(\delta) : \frac{1}{2} (\mathbf{F}^{e\top} \cdot \mathbf{F}^e - \mathbf{I}) = \mathbf{C}(\delta) : \mathbf{E}^e \quad (34)$$

with respect to the intermediate configuration, $F^{e-1}F$. The void volume fraction, ϑ , is a scalar measure of the extent of damage at each point. Here, $C(\vartheta)$ is typically taken to be the damaged elasticity tensor for an isotropic material (Bammann et al., 1993). The Cauchy stress in the current configuration is readily obtained by pushing forward to the current configuration. Provided the applications do not involve extreme hydrostatic pressure, the elastic stretch is negligible compared to unity and we may consider the Cauchy stress, σ , as equivalent to S , apart from a local rigid rotation in $F^e = RU^e$. In this case, we may essentially calculate the rate of change of the Cauchy stress directly in the current configuration, taking appropriate corotational time derivatives to satisfy rotation invariance and accounting for the change of damage through the time step, i.e.,

$$\dot{\sigma} = C(\vartheta) : (D - D^P - D^V - D^{\text{th}}) - \frac{\dot{\vartheta}}{(1-\vartheta)} \sigma = \dot{\sigma} - W^e \cdot \sigma + \sigma \cdot W^e \quad (35)$$

where

$$W^e = W - W^P \quad (36)$$

Here, W is the continuum spin (anti-symmetric part of the velocity gradient) and W^P is the plastic spin, defined as the anti-symmetric part of the velocity gradient associated with F^P , assuming isotropic F^V and F^{th} . The plastic spin essentially reflects effects of texture on the stress-strain response.

The plastic rate of deformation is given by

$$D^P = f(T) G \left(\left\langle \frac{\|s - \alpha\| - (\kappa + Y(T))(1 - \vartheta)}{V(T)(1 - \vartheta)} \right\rangle \right) \frac{s - \alpha}{\|s - \alpha\|} \quad (37)$$

where $(\kappa + Y)(1 - \vartheta)$ is the temperature-dependent shear yield strength of the pure matrix material and $V(T)$ is a temperature-dependent drag strength; κ increases or decreases with cumulative inelastic deformation as a reflection of increased dislocation density. Viscosity function G is a homogeneous function of its argument (cf. Chaboche, 1989), and $\langle \rangle$ are Macauley brackets defined by $\langle P \rangle = P$ for $P \geq 0$; $\langle P \rangle = 0$ otherwise. In equation (37), s is the deviatoric part of the Cauchy stress and α is the deviatoric backstress, due to

dislocation interactions, which resists dislocation motion. To capture dynamic viscoplastic behavior, G is assumed in the Sandia work as a hyperbolic sine form. The plastic rate of deformation D^P includes contribution of only the deviatoric viscoplastic deformation of the metallic matrix. The dilatational component of D is given by

$$D^V = \frac{1}{3} \frac{\dot{\theta}}{1 - \dot{\theta}} I \quad (38)$$

where the void growth rate is based, in the Sandia calculations to be reviewed, on the micromechanical solution of Cocks and Ashby (1980) for intergranular voids growing in a strain hardening, viscous solid, i.e.,

$$\dot{\theta} = \sqrt{\frac{2}{3}} \sinh \left[\frac{2(2n - 1)\sigma_{kt}}{3(2n + 1)\bar{\sigma}} \right] \left(\frac{1}{(1 - \dot{\theta})^n} - (1 - \dot{\theta}) \right) |D^P| \quad (39)$$

where n is considered as a void growth constant. Continuous nucleation of voids is not considered in this particular formulation, but could be included without difficulty (cf. Chu & Needleman, 1980).

The thermal strain rate for cubic polycrystals is given by

$$D^{th} = A \dot{T} I \quad (40)$$

where A is a constant.

Evolution equations are prescribed for the state variables α and κ of the form

$$\dot{\alpha} = h(T) D^P - [r_d(T) |D^P| + r_s(T)] |\alpha| \alpha \quad (41)$$

$$\dot{\kappa} = H(T) |D^P| - [R_d(T) |D^P| + R_s(T)] \kappa^2$$

The superposed circle in equation (41.1) represents the corotational derivative as in equation (10). Temperature-dependent functions r_d and R_d represent dynamic recovery coefficients, while r_s and R_s represent static thermal recovery effects.

In terms of coupling with the adiabatic form of the energy equation, the common simplifying assumption is made that 90% of the plastic work done is dissipated as heat, i.e.,

$$\rho \frac{DE}{Dt} = \rho C_v \dot{T} = 0.9 \sigma : \mathbf{D}^P \quad (42)$$

The failure criterion is idealized as $\vartheta = 1$ or some value close to unity, which is recognized as an approximation. However, since the rate of growth of damage is high for values of damage in excess of 0.1 or so, results are relatively insensitive to the actual value selected.

5.2 Examples

The foregoing model has been implemented in DYNA2D and DYNA3D at Sandia-Livermore, using a modified radial return algorithm, and applied to various nonisothermal problems (cf. Bammann et al., 1993; Lipkin et al., 1988). The code is nearly as efficient as a standard linear hardening elasto-plastic model due to the maintenance of vectorization. Temperature, damage and the ratio of the mean stress to effective stress are assumed constant over the time step, with temperature and damage updated at the end of the time step. Material failure is implemented by progressively permitting the element stiffness to degrade as ϑ accumulates, with a maximum stiffness reduction of 99% to define complete failure. Subsequent to failure, the element mass is maintained, but the element may deform freely in accordance with the motion of adjacent nodes. The plastic spin was neglected in these calculations, as the shear strain (and associated texturing effects) was relatively low.

In this section, we summarize results of predictive calculations for two applications. The first application is subsonic impact of a cylindrical projectile on a circular plate at normal incidence. The second application is that of a focused blast wave interacting with a thin sheet with a hole at both normal and oblique incidence of the blast front.

5.2.1 Penetrator Impact on 6061-T6 Al Plate at Normal Incidence

The first example is the modeling of steel penetrator impacting an aluminum plate at normal incidence (Bammann, 1989; Bammann et al., 1993). The 6061-T6 Al stress-strain response in compression is correlated by the model as shown in Figure 7. This particular material was not significantly strain rate-dependent in the temperature range considered. A

series of circumferentially notched tensile specimens (see Figure 8) were tested quasi-statically at room temperature and atmospheric pressure to estimate the initial value of the porosity, $\vartheta(0)$, in the material as well as the parameter n in the void growth model. In this type of test, a significant hydrostatic tensile stress develops at the center of the specimen, promoting void growth and coalescence at that point; subsequent specimen failure results from progressive outward growth of the porosity, and shear localization effects are effectively avoided except for perhaps the final stage of fracture. One of the four tests was used to fit these parameters and the others are predictions, with excellent agreement in terms of strain to failure over a 25.4 mm gage length spanning the center of the specimen, as shown in Table 3. Note the very significant dependence of ductility on the radius of the notch due to void growth effects.

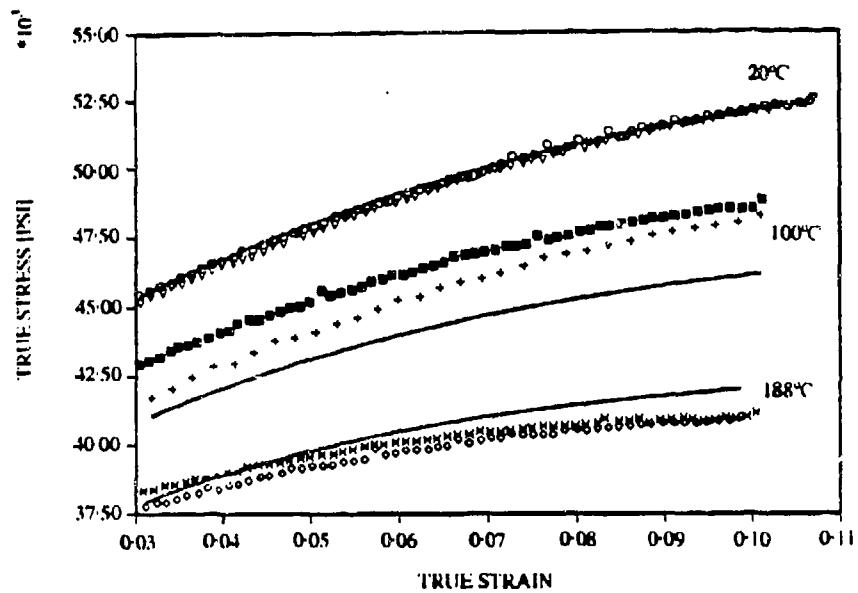


Figure 7. Experimental and Simulated Stress-Strain Response of Al 6061-T6 at Three Temperatures
(Bammann et al., 1993)

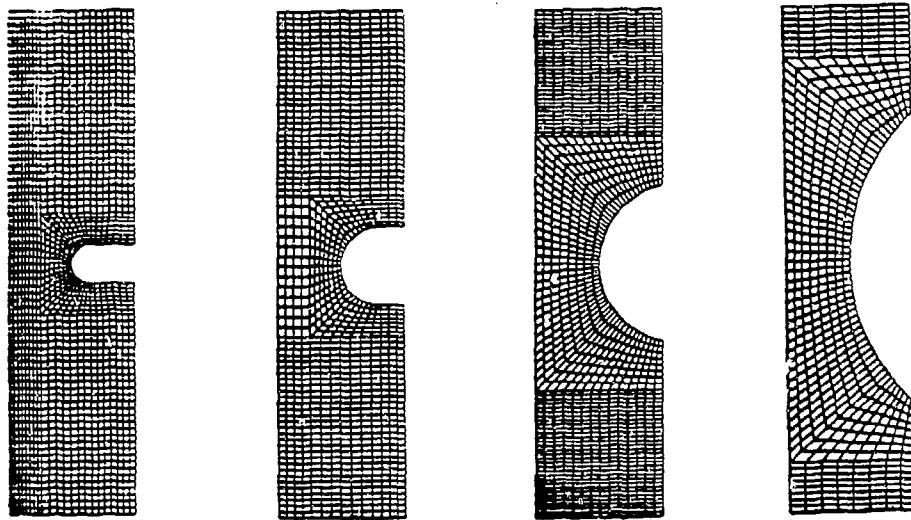
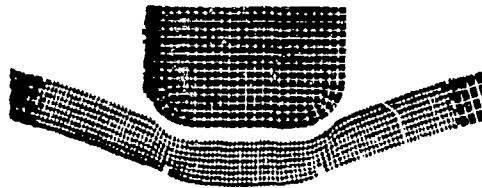


Figure 8. Half Cross-Section of Circumferentially Notched Specimens Used To Calibrate Void Growth Model
 (Bammann et al., 1993)

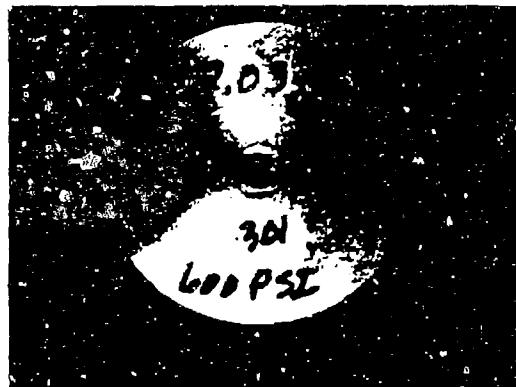
Table 3.
 (Bammann et al., 1993)

Radius (mm)	Number of Tests	Test Average	Calculation
9.9	3	0.043	0.044
4.0	5	0.021	0.023
2.0	5	0.014	0.015
1.0	5	0.011	0.013

Simulations were made of the impact of hardened cylindrical steel rods against 6061-T6 Al disks (diameter of 57.2 mm, thickness of 3.2 mm) with free boundary conditions and compared with experiments. Impact velocities were approximately 92 m/s and 106 m/s, clearly subsonic conditions where through-thickness shear localization and plugging failure might be anticipated as the failure mechanism. Axisymmetric DYNA2D analyses with the internal variable model predicted initial failure (part through cracking) of the plate at impact velocities between 84 m/s and 89 m/s, with a velocity of 102 m/s to 107 m/s required for a plug to be ejected from the plate (complete perforation). Figures 9-10 show the good agreement between the calculations and the experimental results. Even though the local strain rates are higher than those used in characterization experiments, the damage evolution law evidently holds provided the growth of voids is due to matrix plastic deformation.

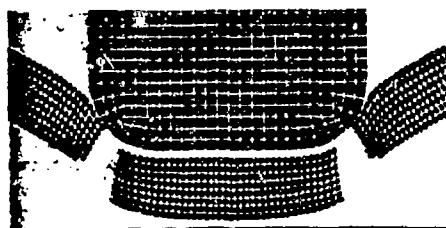


3500 in/s



3610 in/s

Figure 9. Low Velocity Impact and Part Through Cracking
(Bammann et al., 1993)



4200 in/s



4180 in/s

Figure 10. Higher Velocity Impact and Part Through Cracking
(Bammann et al., 1993)

Figures 11-12 compare the effective plastic strain and porosity contours from calculations for an impact velocity of 84 m/s at 200 μ s. The peak value of effective plastic strain is in the interior of the plate and is at a level of 42%, roughly three times higher than the failure strain in a uniaxial test. Hence, an effective plastic strain failure model would have predicted internal failure in the disk but none is observed experimentally at this velocity. The plot of porosity contours in Figure 12 shows a concentration of damage on the back surface of the plate (i.e., incipient plug failure) where it actually does initiate at a

slightly higher velocity, but the level of predicted peak porosity, 5%, is too low for failure to occur. At 89 m/s, however, cracking is predicted to occur by the damage model on the back side of the plate as is experimentally observed at 92 m/s (see Figure 9). Calculations performed with the damage feature inactive revealed that the plastic strain contours differed significantly from those obtained with coupled effects of porosity in terms of location and degree of localization.

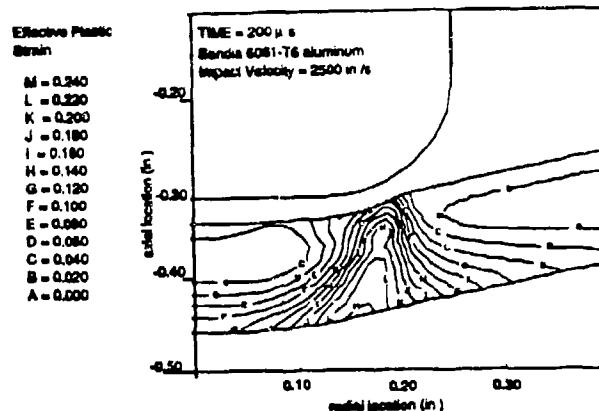


Figure 11. Final Effective Plastic Strain Contours for Impact at 84 m/s (No Failure Occurs) (Bammann et al., 1993)

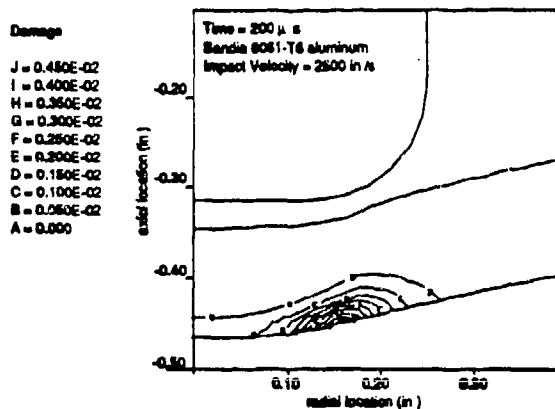


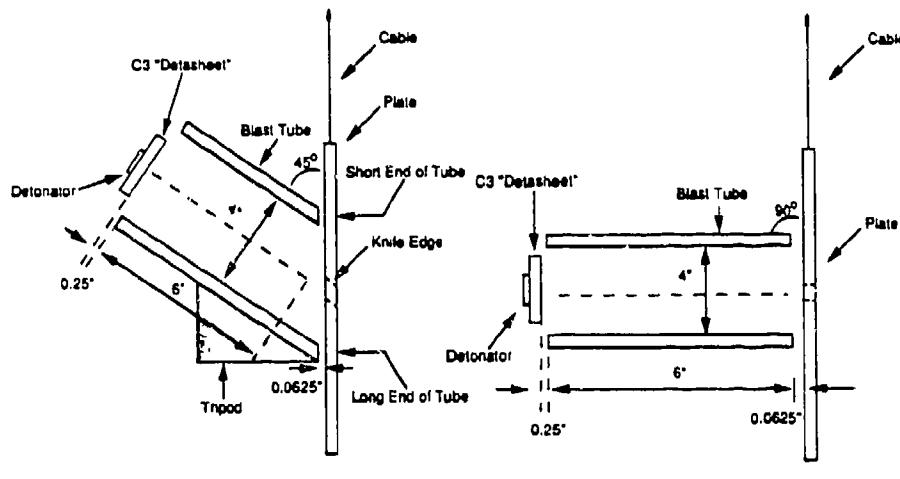
Figure 12. Final Porosity Contours for Impact at 84 m/s (No Failure Occurs) (Bammann et al., 1993)

These analyses largely support the observations and ideas of Cowie et al. (1987) regarding the role of microvoids in shear localization and microcracking. Moreover, they illustrate the important role of accurate, physically based constitutive equations for cases in

which the target is not overmatched by the penetrator for targets of finite thickness. This represents a very large class of practical armor/anti-armor applications.

5.2.2 Focused Blast Waves at Orthogonal and Oblique Incidence

Experiments and DYNA3D solutions for blasts constrained by a steel tube oriented at 45° and 90° to 18" square, 0.0625" thick HY100 and HY130 steel plates with and without central 1" diameter circular holes were conducted and compared for different explosive charges (Bammann et al., 1993). Pressure measurements taken by transducers mounted on a very thick plate mounted at the same location for each level of explosive were matched (in terms of specific impulse) to produce boundary conditions for the DYNA solutions. Figure 13 shows the configuration of the blast tube assembly.



Note: Figures are not to scale

Figure 13. Blast Tube Assembly for Oblique and Orthogonal Blasts
(Bammann et al., 1993)

The DYNA calculations captured not only the details of buckling and overall deformation of the thin plates subjected to blast waves, but also the details of tear initiation sites, tear paths, tear configurations, etc. This is a very significant point. Figure 14 compares the analysis of an orthogonal blast with experimental results. The plate deformations and petaling behavior around the central hole are very faithfully predicted by the analysis. Figure 15 reveals the excellent agreement obtained for a 45° blast at roughly half of the explosive level of the blast in Figure 15. Again, the details of both deformation

and tearing/cracking are accurately described; in this case, initial tearing did not start at the edge of the hole, which was properly predicted.

It is important to note that the tearing and deformation occurs in these experiments in stages or sequences that are not clearly depicted by the final photographs/images in Figures 14-15, but are accurately captured in the numerical simulation.

Dike et al. (1992) have compared the solutions of the Eulerian hydrocode CTH with the experimental measurements of the peak pressure and specific impulse for these focused blast problems. Good agreement was obtained, enabling a link between the CTH calculations of the shock wave in the blast tube and the DYNA3D calculations of the deformation and tearing of the thin plates.

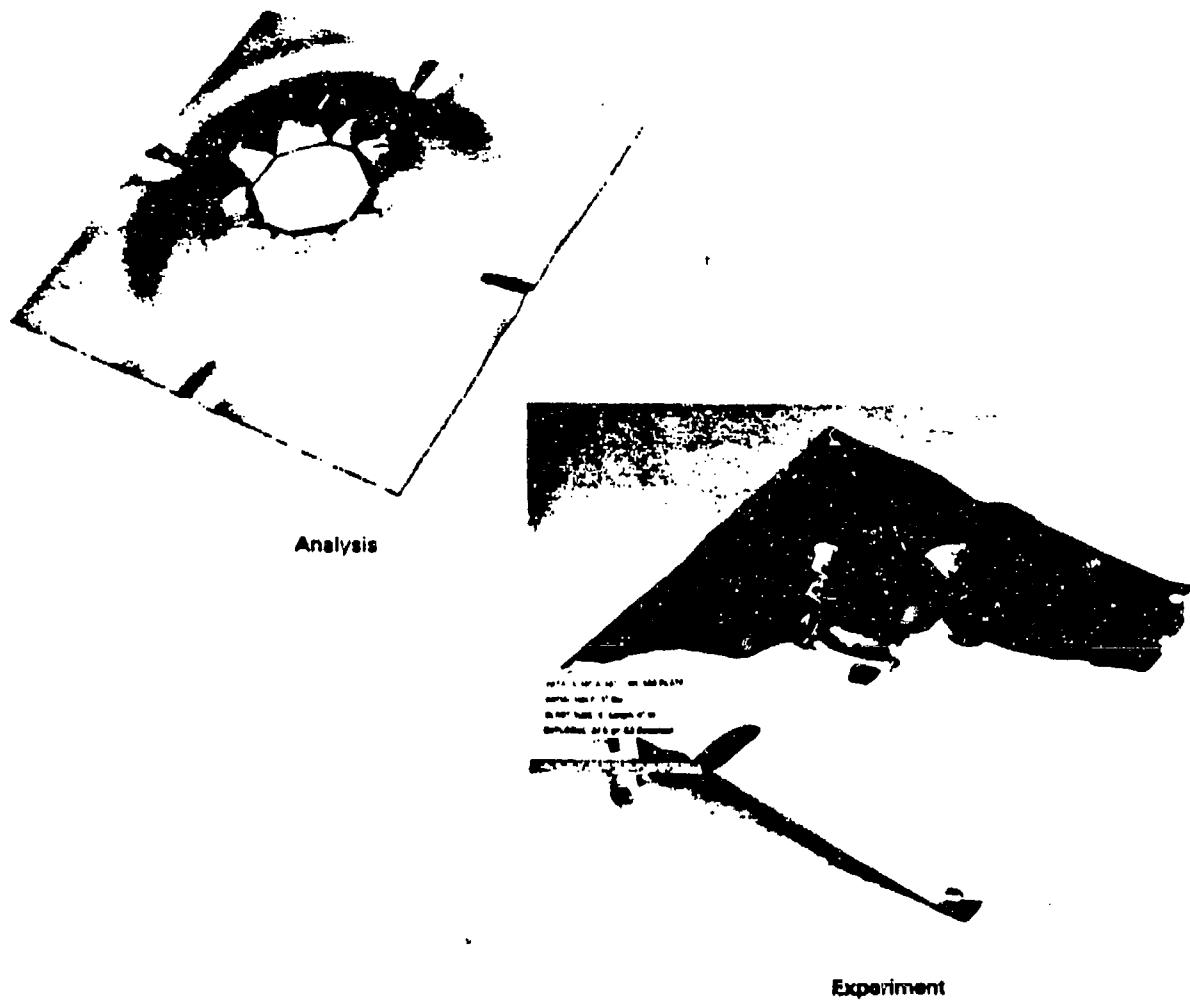
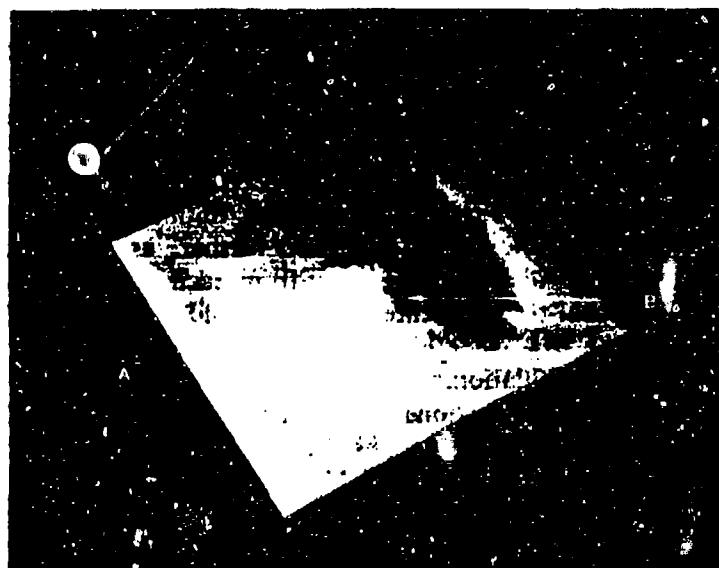
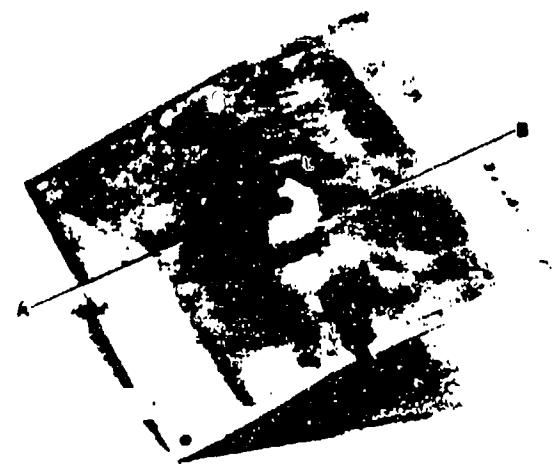


Figure 14. Comparison of Experiment and DYNA3D Simulations of Deformation and Petaling Failure of a Plate with a Central Hole for the Orthogonal Blast (Bammann et al., 1993)



Analysis



Experiment

Figure 15. Comparison of Experiment and DYNA3D Simulations of Deformation and Petaling Failure of a Plate with a Central Hole for the Oblique Blast
(Bammann et al., 1993)

6. CONCLUSIONS

This study has considered several aspects of current technology related to correlation and prediction of damage due to blast waves and ballistic impact, including dynamic material deformation and failure mechanisms, the state-of-the-art in modeling these mechanisms, and implementation of models in hydrocodes. In particular, emphasis is placed on developments in hydrocode modeling and constitutive equations during the past decade. First, we will draw some conclusions regarding recent advances in each of these areas. Then recommendations for future directions will follow.

Models for Dynamic Material Behavior

The understanding of dynamic phenomena has increased significantly due to improved testing and characterization capabilities, along with an emphasis on studying microscale phenomena accompanying material deformation and damage. Significant efforts have been devoted to development of models which describe

- the rate- and temperature-dependence of flow stress,
- the heterogeneity and anisotropy of inelastic deformation,
- void growth and failure by coalescence,
- shear banding,
- microcracking and material erosion,
- fragmentation and spallation, and
- phase changes.

These models have, in principle, supplanted simple failure models based on the instantaneous pressure, maximum tensile stress or the cumulative inelastic strain which were prevalent over a decade ago. In spite of significant advances in micromechanics which address the mechanisms of each of these phenomena, further work is required to develop macroscale constitutive equations which more precisely reflect the physics of each mechanism as well as couplings between them. Moreover, fundamental concepts must be developed for transforming the effects of these mechanisms from the microscale to the macroscale, since macroscale constitutive equations are essential for structural analyses of impact and blast wave problems, even with very substantial increases in computing power.

Recent advances in crystal plasticity, dislocation dynamics, and microfracture have produced promising results. A kinematically and thermodynamically consistent internal

state variable framework for the evolution of deformation and damage at the macroscale is a potentially very promising approach for generalizing microscale models to the scale appropriate for structural calculations. In addition to explicit treatment of damage mechanisms, history dependent processes may be represented by internal state variables. Progress has been made in addressing each of the phenomena listed above using internal state variable theories, but more emphasis is required on the effects of the distribution and interaction at the microscale on the macroscale response.

To accelerate the development of material damage models, as was recommended in the 1980 NMAB report, equal emphasis must be placed on the theoretical development of constitutive laws and on the microstructural characterization which includes the kinetics of transformations, dislocation mechanics, etc. The emerging development of shear band models is an example where idealized computational work has greatly outpaced collaborative efforts from the materials science perspective, resulting in a persistent lack of basic models and computational capability for such localizations.

Development of real-time nondestructive or in-situ capabilities to measure evolving damage would greatly facilitate the identification of couplings between deformation and the evolution of various forms of damage.

Advances in Hydrocodes

The theory of hydrocodes used in explicit calculations of dynamical response during impact of elasto-viscoplastic bodies has been predominately based on the pioneering work of Wilkins (1963). Subsequent generations of hydrocodes have maintained many of the same assumptions. Updated Lagrangian hydrocodes have incorporated more sophisticated treatment of interfaces and mesh rezoning to achieve a description of highly localized interpenetration processes. Eulerian hydrocodes have been developed which preclude rezoning, but face issues associated with convection terms in the time rate of change of stress and internal state variables. Updated Lagrangian hydrocodes perhaps offer the most robust, precise treatment of a wide class of dynamic problems, although the Eulerian hydrocodes may be preferred for certain classes of problems such as penetration in very thick plates. Continued efforts should be devoted to treating the pathologies associated with material convection through the mesh in Eulerian hydrocodes to eventually achieve the same degree of robustness.

Specific items which require significant attention in further development of hydrocodes include:

- Enhancement of modeling algorithms for fragmentation and spallation, including appropriate criteria for failure and treatment of element degradation and momentum exchange of failed elements.
- Documentation and incorporation of micromechanically based evolution laws for damage, shear banding, etc. to advance the basic capabilities of existing hydrocodes.
- Extending Arbitrary Lagrangian Eulerian (ALE) codes to three dimensions, and developing robust adaptive mesh rezoning schemes to account for localization of deformation and damage.

In favor of simplicity, the second point has been largely overlooked in prominent hydrocodes. Relatively dated stress-strain-damage models reflective of the state of understanding in the early to mid-1970's persist in hydrocodes in spite of considerable developments in the past 20 years. It is our impression that this is due to the reluctance of hydrocode developers/users to implement more realistic, albeit complicated material laws. Such laws require more extensive characterization of a material, and knowledge of the physics underlying these models, to determine the constants and to use them effectively. Ultimately, incorporation of these models requires direct interaction of constitutive model developers and hydrocode developers. It is likely necessary for developers of constitutive laws to provide either databases of material properties for a wide range of materials or to provide artificially intelligent, interactive algorithms for determining model constants given mechanical test data. The advantage of this approach is that it yields a predictive impact simulation capability.

There are several promising applications of hydrocodes which may have a strong influence on the use of scale model ballistic testing as well as the development of novel concepts in armor/anti-armor systems:

- Sorting out the influence of size scale on the modeling of penetration, as well as the effects of different phenomena on the penetration process. More realistic treatment of nonlinear wave mechanics for both subsonic and hypervelocity impacts. Development of scaling laws relating laboratory penetration testing to full-scale structural response.
- Development of functionally graded armor, including tailoring of properties of individual layers for specific modes of impact resistance; providing the information necessary to tailor gradients

of initial damage and microstructure in armor which can optimize resistance to penetration, spallation, fragmentation, etc.

An Updated Post-Cold War Research and Development Paradigm

Finally, the post-cold war era offers unique opportunities to alter the research and development paradigm of the past quarter century which essentially consisted of addressing each new development of enemy armor/penetrator capabilities with incremental advances; emphasis was placed on the liberal use of ballistic testing to develop new concepts and verify performance parameters. In many cases, modeling was apparently done as an auxiliary, redundant feature of the research and development effort.

In the post-cold war era, it should be feasible to adopt a longer timeframe for basic research on issues related to armor/anti-armor system development, in lieu of excessive and expensive prototype development. Accordingly, physically based models for deformation and damage mechanisms should be developed along with concurrent improvement of hydrocode capabilities. With this technology, it should eventually become possible to develop new concepts and preliminary designs of armor/anti-armor systems without heavy reliance on testing programs. In addition, the rather extensive technology base developed in these programs should be applied increasingly to problems of enormous industrial significance such as ordinary and explosive forming of sheet materials, explosive processing of metastable materials with novel properties, energy absorption in collisions of transportation vehicles, and protective systems for automotive vehicle occupants, for example.

ACKNOWLEDGMENTS

The authors are grateful to IDA for administrating the fascinating two-year Defense Science Study Group experience which led to this report. The ARPA support of this program is additionally acknowledged. D.L. McDowell appreciates the assistance and cooperation of Sandia National Laboratories in Livermore, CA with regard to the computational examples of an internal state variable theory, principally Dr. D.J. Bammann, Dr. M. Callabresi, M.F. Horstemeyer, and M.L. Chiesa.

REFERENCES

- _____, 1980, "Materials Response to Ultra-High Loading Rates," National Materials Advisory Board, NMAB-356, National Academy of Sciences, Wash., D.C.
- Aifantis, E.C., 1986, "On the Structure of Single Slip and its Implications for Inelasticity, in Large Deformation of Solids, (Gittus, Zarka & Nemat-Nasser, eds.), Chap. 17, Elsevier.
- Aifantis, E.C., 1987, "The Physics of Plastic Deformation," Int. J. Plasticity, Vol. 3, pp. 211-247.
- Allen, D.H., 1991, "Thermomechanical Coupling in Inelastic Solids," Applied Mechanics Reviews, Vol. 44, No. 8, pp. 361-373.
- Anderson, C.E., Jr., 1987, "An Overview of the Theory of Hydrocodes," Int. J. Impact Engineering, 5, p. 33.
- Anderson, C.E., Jr. and Bodner, S.R., 1988, "Ballistic Impact: The Status of Analytical and Numerical Modeling," Int. J. Impact Engineering, Vol. 7, No. 1, pp. 9-35.
- Anderson, C.E., Jr., Mullin, S.A. and Kuhlman, C.J., 1992, "Strain-Rate Effects in Replica Scale Model Penetration Experiments," SwRI Report 3593/002, August.
- Andersson, H., 1977, "Analysis of a Model for Void Growth and Coalescence Ahead of a Moving Crack Tip," J. Mech. Phys. Solids, Vol. 25, pp. 217-233.
- Aravas, N. and E.C. Aifantis, 1991, "On the Geometry of Slip and Spin in Finite Plastic Deformation," Int. J. Plasticity, Vol. 7, pp. 141-160.
- Asaro, R.J., 1983, "Crystal Plasticity," ASME J. Appl. Mech., 50, pp. 921-934.
- Asaro, R.J. and Lowe, T.C., 1985, "Crystal Plasticity Analysis of Large-Strain Shear," Sandia Report 85-8245, Livermore, CA.
- Asay, J.R. and Kerley, G.I., 1986, "The Response of Materials to Dynamic Loading," Int. J. Impact Engineering, Vol. 5.
- Backman, M.E. and Finnegan, S.A., 1985, "Experimental Data for Characterizing Perforating Impacts: Fragmentation Processes," NWC TP 6490, Naval Weapons Center, China Lake, CA, May.
- Bammann, D.J., 1984, "An Internal Variable Model of Viscoplasticity," in Media with Microstructures and Wave Propagation, Eds. Aifantis, E.C. and Davison, L., Int. J. Engineering Science, Vol. 8-10, Pergamon Press, p. 1041.
- Bammann, D.J., 1985, "An Internal Variable Model of Elastic-Viscoplasticity," in The Mechanics of Dislocations, Eds. Aifantis, E.C. and Hirth, J.P., American Society of Metals: 203.
- Bammann, D.J., and Johnson, G.C., 1987, "On the Kinematics of Finite-Deformation Plasticity," Acta Mech., Vol. 70, pp. 1-13.
- Bammann, D.J. and Aifantis, E.C., 1987, "A Model for Finite-Deformation Plasticity," Acta Mech., Vol. 69, pp. 97-117.

- Bammann, D.J., 1989, "An Anisotropic Hardening Model of Plasticity," IDA Document D-576, Proc. DoD/DARPA Coordination Mtg. on Advanced Armor/Anti-Armor Materials and Advanced Computational Methods, Ed. G. Mayer, Washington, D.C., Dec. 13-15, 1988.
- Bammann, D.J., 1990, "Modeling the Temperature and Strain Rate Dependent Large Deformation of Metals," Applied Mechanics Reviews, Vol. 43, No. 5, Part 2, E. Krempl and D.L. McDowell, eds., pp. S312-S319.
- Bammann, D.J., Chiesa, M.L., Horstemeyer, M.F., and Weirgarten, L.I., 1993, "Failure in Ductile Materials Using Finite Element Methods," Chapter 1, Structural Crashworthiness and Failure, Ed. N. Jones and T. Wierzbicki, Elsevier Appl. Sci. Publ, New York, pp. 1-54.
- Bassani, J.L., 1990, "Latent Hardening and Single Crystal Plasticity," Applied Mechanics Reviews, Vol. 43, No. 5, Part 2, E. Krempl and D.L. McDowell, eds., p. S320.
- Batra, R.C., Chen, X.J., and Peng, Z., 1993, "Axisymmetric Penetration of Thermoviscoplastic Targets," Proc. U.S. Army Symp. on Solid Mechanics, 17-19 August, Plymouth, MA.
- Bay, B., Hansen, N. and Kuhlmann-Wilsdorf, D., 1989, "Deformation Structures in Lightly Rolled Pure Aluminum," Materials Science and Engineering, A113, pp. 385-397.
- Bay, B., Hansen, N., Hughes, D.A. and Kuhlmann-Wilsdorf, D., 1992, "Evolution of F.C.C. Deformation Structures in Polyslip," Acta Metall. Mater., Vol. 40, No. 2, pp. 205-219.
- Becker, R.C. and A. Needleman, 1986, "Effect of Yield Surface Curvature on Necking and Failure in Porous Plastic Solids," ASME J. Applied Mech., Vol. 53, pp. 491-499.
- Becker, R.C., Smelser, R.E. and Richmond, O., 1989, "The Effect of Void Shape on the Development of Damage and Fracture in Plane-Strain Tension," J. Mech. Phys. Solids, Vol. 37, p. 515.
- Belytschko, T., Liu, W.K., Kennedy, J.M. and Ong, S.J., 1984, "Hourglass Control in Linear and Nonlinear Problems," Comp. Meth. in Appl. Mech. Engineering, Vol. 43, pp. 251-276.
- Bodner, S. and Partom, Y., 1975, "Constitutive Equations for Elastic-Viscoplastic Strain Hardening Materials," ASME J. Appl. Mech., pp. 385-389.
- Chaboche, J.L., 1989, "Constitutive Equations for Cyclic Plasticity and Cyclic Viscoplasticity," International Journal of Plasticity, Vol. 5, No. 3, p. 247.
- Charter, A.C. and Orphal, D.L., 1991, "The Segmented Rod: A New Penetrator Concept for the Defeat of Advanced Armor," Journal of Defense Research, Vol. 20, No. 2, p. 531 (U).
- Chen, Z., 1991, "Experimental, Theoretical and Computational Investigation of Failure of Quasi-Brittle Structures," Final Report, Phillips Lab., Air Force Systems Command, AD-B158 333.
- Chhabildas, L.C. and Asay, J.R., 1978, "Shock and Unloading Wave Profiles in Metals at High Pressure," Bull. Am. Phys. Soc., Vol. 23, p. 69.
- Chu, C.C. and Needleman, A., 1980, "Void Nucleation Effects in Biaxially Stretched Sheets," J. Engineering Mater. Techn., Vol. 102, pp. 249-256.

- Cleja-Tigoiu, S. and Soos, E., 1990, "Elastoviscoplastic Models with Relaxed Configurations and Internal State Variables," Applied Mechanics Reviews, Vol. 43, No.7, July, pp. 131-151.
- Cocks, A.C.F. and Ashby, M.G., 1980, "Intergranular Fracture During Power Law Creep Under Multiaxial Stresses," Metal Science, Aug.-Sept., pp. 395-402.
- Cocks, A.C.F., 1989, "Inelastic Deformation of Porous Materials," J. Mech. Phys. Solids, Vol. 37, No. 6, pp. 693-715.
- Costin, L.S., 1983, "A Microcrack Damage Model for Brittle Rock," Sandia Report SAND83-1590.
- Cowie, J.G., Azrin, M. and Olson, G.B., 1987, "Microvoid Formation During Shear Deformation of Ultrahigh Strength Steels," in Innovations in High Strength Steels, Proc. 34th Sagamore Conf., U.S. Army, p. 357.
- Cuitino, A.M. and Ortiz, M., 1992, "Computational Modeling of Single Crystals," Modelling Simul. Mater. Sci. Eng., Vol. 1, pp. 225-263.
- Curran, D.R. Seaman, L. and Shockley, D.A., "Dynamic Failure in Solids," Physics Today, January 1977, pp. 46-55.
- Dafalias, Y.F., 1983, "Corotational Rates for Kinematic Hardening at Large Plastic Deformations," ASME Journal of Applied Mechanics, Vol. 50, pp. 561-565.
- Dafalias, Y.F., 1985, "The Plastic Spin," ASME Journal of Applied Mechanics, Vol. 107, pp. 865-871.
- Dafalias, Y.F., 1987, "Issues on the Constitutive Formulation at Large Elastoplastic Deformations, Part 1: Kinematics," Acta Mechanica, Vol. 69, pp. 119-138.
- Dehn, J.T., 1986, "A Unified Theory of Penetration," BRL-TR-2770, Dec.
- Dienes, J.K., 1984, "A Statistical Theory of Fragmentation Processes," LA-UR-84-3173, Los Alamos, NM.
- Dike, J.J., Horstemeyer, M.F. and Weingarten, L.I., 1992, "Modeling Focused Blasts with a Hydrodynamic Code," 63rd Shock and Vibration Symp., 27-29 Oct., Las Cruces, NM, p. 507.
- Dike, J.J., Jin, P.S., Revelli, V.D. and Trento, W.P., 1993, "Linking a Hydrodynamic Code and a Finite Element Code to Predict Underwater Explosion Damage," SAND 92-8619, UC-705, Feb.
- Doraivelu, S.M., Gegel, H.L., Gunasekera, J.S., Malas, J.C., Thomas, J.F. and Morgan, J.T., 1984, "A New Yield Function for Compressible P/M Materials," International Journal of Mechanical Science, Vol. 26, pp. 527-535.
- Dragon, A. and Chihab, A., 1985, "On Finite Damage: Ductile Fracture-Damage Evolution," Mechanics of Materials, Vol. 4, pp. 95-106.
- Drucker, D.C., 1949, "Relation of Experiments to Mathematical Theories of Plasticity," ASME Journal of Applied Mechanics, Vol. 16, pp. 349-357.
- Eftis, J. and Nemes, J.A., 1991, "Evolution Equation for the Void Volume Growth Rate in a Viscoplastic-Damage Constitutive Model," Int. J. Plasticity, Vol. 7, pp. 275-293.
- Erlich, D.C., Seaman, L., Shockley, D.A. and Curran, D.R., 1980, "Development and Application of a Computational Shear Band Model," Contract Report No. ARBRL-CR-00416, Ballistic Research Laboratory, Aberdeen, MD, March.

- Ferrari, G., 1988, "The "Hows" and "Whys" of Armour Penetration," MILTECH, Oct. 1988, pp. 81-96.
- Follansbee, P.S. and Kocks, U.F., 1988, "A Constitutive Description of the Deformation of Copper Based on the Use of the Mechanical Threshold Stress as an Internal State Variable", *Acta Metallurgica*, Vol. 36, No. 1, pp. 81-93.
- Follansbee, P.S., 1989, "Dynamic Deformation and Fracture in Armor/Anti-Armor Materials," IDA Document D-576, Proc. DoD/DARPA Coordination Mtg. on Advanced Armor/Anti-Armor Materials and Advanced Computational Methods, Ed. G. Mayer, Washington, D.C., Dec. 13-15, 1988.
- Garrison, W.M., Jr. and Handerhan, K.J., 1987, "Fracture-Toughness: Particle-Dispersion Correlations," in Innovations in High Strength Steels, Proc. 34th Sagamore Conf., U.S. Army, p. 443.
- Germain, P., Nguyen, Q.S. and Suquet, P., 1983, "Continuum Thermodynamics," ASME J. Appl. Mech., Vol. 50, p. 1010.
- Goel, R.A., Chandra, S., Abrol, A.K. and Singh, B., 1988, "Behaviour of a Kinetic Energy Projectile on Angular Impact," Def. Sci. J., Vol. 38, No. 3, July, pp. 293-299.
- Goodman, H.J., 1982, "Measures of Explosive Effectiveness and Effects of Charge Geometry," Tech. Report ARBRL-TR-02434, Oct.
- Grady, D.E., 1982, "Local Inertial Effects in Dynamic Fragmentation," J. Appl. Phys., Vol. 53, pp. 322-325.
- Grady, D.E. and Kipp, M.E., 1985, J. Appl. Phys., Vol. 58, p. 1221.
- Grady, D.E. and Kipp, M.E., 1989, "Fragmentation of Solids Under Dynamic Loading," in Structural Failure, John Wiley & Sons.
- Grady, D.E., 1990, J. Appl. Phys., Vol. 68, No. 12, pp. 6099-6105.
- Gupta, N.K. and Madhu, V., 1992, "Normal and Oblique Impact of a Kinetic Energy Projectile on Mild Steel Plates," Int. J. Impact Engineering, Vol. 12, No 3, pp. 333-343.
- Gurson, A.L., 1977, "Continuum Theory of Ductile Rupture by Void Nucleation and Growth: Part I - Yield Criteria and Flow Rules for Porous Ductile Media," ASME J. Engineering Mater. Techn., Vol. 99, pp. 1-15.
- Hallquist, J.O., 1988, "User's Manual for DYNA-2D - An Explicit Two-Dimensional Hydrodynamic Finite Element Code with Interactive Rezoning and Graphical Display," UCID - 18756, Rev. 3, Lawrence Livermore National Laboratory, Livermore, CA, March.
- Hertel, E.S., Jr., 1992, "A Comparison of the CTH Hydrodynamics Code with Experimental Data," SAND92-1879, UC-410, Sept.
- Hornemann, U., Schroder, G.A., and Weimann, K., 1987, "Explosively-formed Projectile Warheads," Military Technology, April, pp. 36-51.
- Huffington, N.J., 1991, "A Reexamination of the Plastic Flow Criterion for Copper," Proc. 12th Army Symposium on Solid Mechanics, Ed. S.C. Chou, Plymouth, MA, pp. 509-520.
- Hughes, D.A. and Nix, W.D., 1989, Materials Science and Engineering, A122, pp. 153-172.

- Hughes, D.A. and Hansen, N., 1991, Materials Science and Technology, 7, pp. 544-553.
- Hutchinson, J. and Tvergaard, V., 1987, "Microvoid Nucleation Softening in Metals," Proc. 34th Sagamore Army Materials Research Conf.
- Johnson, W.E. and Anderson, C.E., Jr., 1987, "History and Application of Hydrocodes in Hypervelocity Impact," Int. J. Impact Engineering, Vol. 5, pp. 423-439.
- Johnson, G.R. and Cook, W.H., 1983, "A Constitutive Model and Data for Metals Subjected to Large Strains, High Strain Rates and High Temperatures," Proc. Seventh Int. Symp. on Ballistics.
- Johnson, G.R. and Cook, W.H., 1985, "Fracture Characteristics of Three Metals Subjected to Various Strains, Strain-Rates, Temperatures and Pressures," J. Engineering Frac. Mech., Vol. 21, No. 1, pp. 31-48.
- Johnson, G.R. and Stryk, R.A., 1987, "Eroding Interface and Improved Tetrahedral Element Algorithms for High-Velocity Impact Computations in Three Dimensions," Int. J. Impact Engineering, Vol. 5, pp. 411-421.
- Johnson, G.R. and Holmquist, T.J., 1989, "Test Data and Computational Strength and Fracture Model Constants for 23 Materials Subjected to Large Strains, High Strain Rates and High Temperatures," LANL Report LA-11463-MS, January.
- Jones, N., 1989, "Recent Studies on the Dynamic Plastic Behavior of Structures," Applied Mechanics Reviews, Vol. 42, No. 4, pp. 95-115.
- Kim, K.T. and Carroll, M.M., 1987, "Compaction Equations for Strain-Hardening Porous Materials," Int. J. Plasticity, Vol. 3, pp. 63-73.
- Kimsey, K.D. and Zukas, J.A., 1986, "Contact Surface Erosion for Hypervelocity Problems," BRL-MR-3495, U.S. Army BRL.
- Kipp, M.E., Grady, D.E. and Chen, E.P., 1980, "Strain-Rate Dependent Fracture Initiation," Int. J. Fracture, Vol. 16, pp. 471-478.
- Klopp, R.W. and Shockley, D.A., 1992, "Tests for Determining Failure Criteria of Ceramics Under Ballistic Impact," Final Report, U.S. Army Research Office, Contract No. DAAL03-88-K-0200.
- Korjack, T.A., 1992, "A System Structure for Predictive Relations in Penetration Mechanics," BRL Tech. Report BRL-TR-3317, Feb.
- Krajcinovic, D. and Fonseka, G.U., 1981, "The Continuous Theory of Brittle Materials, Part 1: General Theory," ASME J. Appl. Mech., Vol. 48, pp. 809-815.
- Krause, C.D. and Raftenberg, M.N., 1993, "Metallographic Observations of Rolled-Homogeneous Armor Specimens from Plates Perforated by Shaped Charge Jets," Army Research Laboratory, ARL-MR-68, 1993.
- Lee, Y.K., 1988, "A Finite Elastoplastic Flow Theory for Porous Media," Int. J. Plasticity, Vol. 4, pp. 301-316.
- Lemaitre, J., and Chaboche, J.L., 1985, Mecanique des Materiaux Solides, Dunod Publ., Paris.
- Lipkin, J., Chiesa, M.L. and Bammann, D.J., 1989, "Thermal Softening of 304L Stainless Steel: Experimental Results and Numerical Simulations," Proc. Int. Conf. on Impact Loading Dynamic Behavior of Materials, Eds. Chiem, C.Y., Kunze, H.D. and Meyer, L.W., Fraunhofer Institute, Bremen, pp. 687-694.

- Mandel, J., 1971, "Plasticite Classique et Viscoplasticite (courses and lectures, No. 97)," CISM. New York: Udine/Springer.
- Mandel, J., 1973, "Equations Constitutives et Directeurs dans les Milieux Plastiques et Viscoplastiques," Int. J. Sol. Struct., Vol. 9, pp. 725-740.
- Mandel, J. 1974, "Thermodynamics and Plasticity," in Foundations of Continuum Thermodynamics, J.J. Delgado et al., Eds. New York: Macmillan.
- Mandell, D.A. and Henninger, R., 1992, "Ceramics Models in the MESA Codes," 1992 Hypervelocity Impact Symp., Los Alamos National Laboratory, LA-UR-92-3003.
- Mandell, D.A., 1993a, "Prediction of Alumina Penetration," LA-12520, UC-700.
- Mandell, D.A., 1993b, "Nonlinear Wave Predictions in Ceramics," IUTAM Symp., Aug. 15-20, University of Victoria, Victoria, BC, Canada, LA-UR-93-2437.
- Margolin, L.G., 1983, "Elastic Moduli of a Cracked Body," Int. J. Fracture, Vol. 22, 1983, pp. 65-79.
- Margolin, L.G., 1984, "Microphysical Models for Inelastic Material Response," Int. J. Engineering Sci., Vol. 22, No. 8-10, pp. 1171-1179.
- Mayer, G., 1992, "New Directions in Research on Dynamic Deformation of Materials," IDA Paper P-2499 (Revised), Feb.
- McClintock, F.A., 1968, "A Criterion for Ductile Fracture by the Growth of Holes," ASME J. Appl. Mech., Vol. 35, p. 363.
- McDowell, D.L., 1985, "An Experimental Study of the Structure of Constitutive Equations for Nonproportional Cyclic Plasticity," ASME Journal of Engineering Materials and Technology, Vol. 107, pp. 307-315.
- McDowell, D.L., Stahl, D.R., Stock, S.R. and Antolovich, S.D., 1988, "Biaxial Path Dependence of Deformation Substructure of Type 304 Stainless Steel," Met. Trans. A, 19A, pp. 1277-1293.
- McDowell, D.L., Miller, M.P., and Bammann, D.J., 1991, "Some Additional Considerations for Coupling of Material and Geometric Nonlinearities for Polycrystalline Metals," MECAMAT '91 Proceedings of the International Seminar on Large Plastic Deformations, Fountainebleau, France.
- McDowell, D.L. and Moosbrugger, J.C., 1992, "Continuum Slip Foundations of Elasto-Viscoplasticity," Acta Mechanica, Vol. 93, pp. 73-87.
- McDowell, D.L., 1992, "A Nonlinear Kinematic Hardening Theory for Cyclic Thermoplasticity and Thermoviscoplasticity," Int. J. Plasticity, Vol. 8, pp. 695-728.
- McGlaun, J.M., 1992, "CTH Reference Manual: Cell Thermodynamics," SAND91-0002, UC-410.
- McGlaun, J.M. and Thompson, S.L., 1990, "CTH: A Three-Dimensional Shock Wave Physics Code," Int. J. Impact Engineering, Vol. 10., pp. 351-360.
- Mear, M.E. and Hutchinson, J.W., 1985, "Influence of Yield Surface Curvature on Flow Localization in Dilatant Plasticity," Mechanics of Materials, Vol. 4, p. 395.
- Mear, M.E., 1990, "On the Plastic Yielding of Porous Metals," Mech. Materials, Vol. 9, pp. 33-48.

- Mecking, H., and Kocks, U.F., 1981, "Kinetics of Flow and Strain Hardening," *Acta Metallurgica*, Vol. 29, pp. 1865-1875.
- Mescall, J.F. and Rogers, H., 1987, "The Role of Shear Instability in Ballistic Penetration," in *Innovations in High Strength Steels*, Proc. 34th Sagamore Conf., U.S. Army, p. 287.
- Meyers, M.A. and Murr, L.E., 1980, "Defect Generation in Shock-Wave Deformation," in *Shock Waves and High-Strain-Rate Phenomena in Metals*, Eds. M.A. Meyers and L.E. Murr, Int. Conf. on Metallurgical Effects of High-Strain-Rate Deformation and Fabrication, Albuquerque, June 22-26, pp. 487-530.
- Miller, A.K., Ed., 1987, Unified Constitutive Equations for Creep and Plasticity, Elsevier Appl. Sci., New York.
- Miller, M.P. and McDowell, D.L., 1992, "Stress State Dependence of Finite Strain Inelasticity," *Microstructural Characterization in Constitutive Modeling of Metals and Granular Media*, ASME MD - Vol. 32, Ed. G.Z. Voyatzis, Phoenix, AZ, April, pp. 27-44.
- Mock, W., Jr. and Holt, W.H., 1983, "Fragmentation Behavior of Armco iron and HF-1 Steel Explosive-Filled Cylinders," *J. Appl. Physics*, Vol. 54, No. 5, pp. 2344-2351.
- Molinari, A. and Clifton, R.J., 1987, "Analytical Characterization of Shear Localization in Thermoviscoplastic Materials," *ASME J. Appl. Mech.*, Vol. 54, pp. 806-812.
- Moosbrugger, J.C. and McDowell, D.L., 1990, "A Rate Dependent Bounding Surface Model with a Generalized Image Point for Cyclic Nonproportional Viscoplasticity," *Journal of Mechanics and Physics of Solids*, Vol. 38, No. 5, pp. 627-656.
- Mueller, A.C., 1992, "Microfracturing-Based Continuum Damage Models for Hydrocode Applications," Phase I Final Report, Contract No. DAAH01-92-C-R165, DARPA.
- Narasimhan, R., Rosakis, A.J. and Moran, B., 1992, "A Three Dimensional Numerical Investigation of Fracture Initiation by Ductile Failure Mechanisms in a 4340 Steel," in *Advances in Fracture/Damage Models for the Analysis of Engineering Problems*, ASME AMD-137, pp. 13-53.
- Needleman, A. and Rice, J.R., 1978, "Limits to Ductility Set by Plastic Flow Localization," in Mechanics of Sheet Metal Forming, eds. D.P. Koistinen and N.M. Wang, Plenum Press, New York, pp. 237-267.
- Needleman, A. and Tvergaard, V., 1987, "An Analysis of Ductile Rupture Modes at a Crack Tip," *J. Mech. Phys. Solids*, Vol. 35, No. 2, pp. 151-183.
- Olson, G.B., Mescall, J.F. and Azrin, M., 1980, "Adiabatic Deformation and Strain Localization," Proc. Int. Conf. Metallurgical Effects of High Strain-Rate Deformation and Fabrication, Albuquerque, Plenum Press.
- Olson, G.B., Anctil, A.A., DeSisto, T.S., and Kula, E.B., 1983, "Anisotropic Embrittlement in High-Hardness ESR 4340 Steel forgings," *Metall. Trans.*, Vol. 14A, p. 1661.
- Oyane, M., Shima, S. and Kono, Y., 1973, "Theory of Plasticity for Porous Metals," *Bulletin of the JSME*, Vol. 16, No. 99, pp. 1254-1262.
- Perzyna, P., 1984, "Constitutive Modeling of Dissipative Solids for Postcritical Behavior and Fracture," *ASME J. Engineering Mater. Techn.*, Vol. 106, pp 410-419.

- Predebon, W.W., Anderson, C.E., Jr. and Walker, J.D., 1991, "Inclusion of Evolutionary Damage Measures in Eulerian Wavecodes," Computational Mechanics, Vol. 7, pp. 221-236.
- Qiu, Y.P. and Weng, G.J., 1992, "An Energy Approach to the Plasticity of Porous Materials," ASME AMD-Vol. 132/MD-Vol. 30, in Recent Advances in Damage Mechanics and Plasticity, ed. J.W. Ju, pp. 203-217.
- Quidot, M., 1987, "Dynamic Fragmentation of Compact Energetic Materials," Int. Conf. on Impact Loading and Dynamic Behaviour of Materials, Bremen, Germany, pp. 609-614.
- Radin, J. and Goldsmith, W., 1988, "Normal Projectile Penetration and Perforation of Layered Targets," Int. J. Impact Engineering, Vol. 7, No. 2, pp. 229-259.
- Raftenberg, M.N., 1992a, "Modeling RHA Plate Perforation by a Shaped Charge Jet," BRL Tech. Report BRL-TR-3363.
- Raftenberg, M.N., 1992b, "Experimental Investigation of RHA Plate Perforation by a Shaped-Charge Jet," Proc. Army Symp. on Solid Mechanics, Ed. S.C. Chou, U.S. ARMY Materials Techn. Lab., Watertown, MA, pp. 395-410.
- Rajendran, A.M. and Cook, W.H., 1988, "A Comprehensive Review of Modeling of Impact Damage in Ceramics," Final Report, Air Force Armament Laboratory, AFATL-TR-88-143.
- Rhode, R.W., 1970, "Temperature Dependence of the Shock-Induced Reversal of Martensite to Austenite in an Iron-Nickel-Carbon Alloy," Acta Metall., Vol. 18, p. 903.
- Rice, J.R. and Tracey, D.M., 1969, "On the Ductile Enlargement of Voids in Triaxial Stress Fields," J. Mech. Phys. Solids, Vol. 17, p. 201.
- Rice, J.R. 1971, "Inelastic constitutive relations for solids: an internal variable theory and its application to metal plasticity", J. Mech. Phys. Solids 19: 433-455.
- Rice, J.R. and Johnson, M.A., 1970, "The Role of Large Crack Tip Geometry Changes in Plane Strain Fracture," in Inelastic Behavior of Solids, Ed. M.F. Kanninen et al., McGraw-Hill.
- Rogers, H.C. and Shastry, C.V., 1980, "Material Factors in Adiabatic Shearing in Steels," Proc. Int. Conf. on Metallurgical Effects of High Strain Rate Deformation and Fabrication," Albuquerque, NM, Plenum, p. 285.
- Rollett, A.D., Jensen, D.J., and Stout, M.G., 1992, "Modelling the Effect of Microstructure on Yield Anisotropy," Proc. 13th Risø Int. Symp. on Materials Science, Eds. S.I. Andersen et al., Risø National Laboratory, Roskilde, Denmark, pp. 93-109.
- Rousselier, G., 1981, "Finite Deformation Constitutive Relations Including Ductile Fracture Damage," in Three Dimensional Constitutive Relations and Ductile Fracture, Ed. S. Nemat-Nasser, North Holland Publ., pp. 331-355.
- Scheidler, M., 1993, "On the Coupling of the Pressure and the Deviatoric Stress in Hyperelastic Materials," Proc. U.S. Army Symp. on Solid Mechanics, 17-19 August, Plymouth, MA.
- Seaman, L., Curran, D.R. and Murri, W.J., 1985, "A Continuum Model for Dynamic Tensile Microfracture and Fragmentation," ASME J. Appl. Mech., Vol 52, pp. 593-600.

- Senior, B.A., Noble, F.W. and Eyre, B.L., 1986, "The Nucleation and Growth of Voids at Carbides in 9 Cr-1 Mo Steel," *Acta Metall.*, Vol. 34, p. 1321.
- Shockey, D.A., Curran, D.R. and DeCarli, P.S., 1975, *J. Appl. Phys.*, Vol. 46, p. 3766.
- Shockey, D.A. and Erlich, D.C., 1980, "Metallurgical Influences on Shear Band Activity," *Proc. Int. Conf. Metallurgical Effects on High Strain-Rate Deformation and Fabrication*, Albuquerque, Plenum Press, p. 249.
- Shockey, D.A., Marchand, A.H., Skaggs, S.R., Cort, G.E., Burkett, M.W. and Parker, R., 1990, "Failure Phenomenology of Confined Ceramic Targets and Impacting Rods," *Int. J. Impact Engineering*, Vol. 9, No. 3, pp. 263-275.
- Sidoroff, F. and Teodosiu, C., 1986, Microstructure and Phenomenological Models for Metals. Large Deformation of Solids. Physical Basis and Mathematical Modelling, Eds. J. Gittus et al., Elsevier Appl. Sci. Publ., New York, pp. 163-186.
- Silling, S.A., 1991, "CTH Reference Manual: Viscoplastic Models," SAND91-0292, UC-405.
- Silling, S.A., 1992, "Eulerian Simulation of the Perforation of Aluminum Plates by Nondeforming Projectiles," SAND92-0493, UC-405.
- Smith, D.L., 1983, "A Subroutine to Isolate Material Packages in the HELP Hydrodynamic Code," Department of Defence, Report MRL-R-872, Melbourne, Victoria.
- Sternberg, J., 1985, "A Critical Review of Target Strength Effects in Resisting High Velocity Penetration," RDA-TR-188900-001, DARPA report, October.
- Sun, D.-Z., Siegelse, D., Voss, B. and Schmitt, W., 1989, "Application of Local Damage Models to the Numerical Analysis of Ductile Rupture," *Fatigue Fract. Engineering Mater. Struct.*, Vol. 12, No. 3, pp. 201-212.
- Swegle, J.W. and Grady, D.E., 1985, "Shock Viscosity and the Prediction of Shock Wave Rise Times," *J. Appl. Phys.*, Vol. 58, No. 2, pp. 692-701.
- Tate, A., 1967, "A Theory for the Deceleration of Long Rods After Impact," *J. Mech. Phys. Solids*, Vol. 15, pp. 387-399.
- Tate, A., 1969, "Further Results in the Theory of Long Rod Penetration," *J. Mech. Phys. Solids*, Vol. 17, pp. 141-150.
- Taylor, P.A., 1992, "CTH Reference Manual: The Steinberg-Guinan-Lund Viscoplastic Model," SAND92-0716-UC-405.
- Taylor, P.A. and Dodson, B.W., 1985, "Simulation of Lattice Damage due to Dynamic Loading," SAND-85-0391C.
- Teodosiu, C., 1991, "Texture Vs. Microstructure in Anisotropic Plasticity," in Anisotropy and Localization of Plastic Deformation, J.-P. Boehler and A.S. Khan eds, pp. 179-182.
- Trucano, T.G., 1993, "Equation of State and Fragmentation Issues in Computational Lethality Analysis," Sandia Report SAND92-2397, UC-905.
- Tuler, F.R. and Butcher, B.M., 1968, "A Criterion for the Time Dependence of Dynamic Fracture," *Int. J. Frac. Mech.*, Vol. 4, No. 4, pp. 431-437.
- Tvergaard, V., 1981, "Influence of Voids on Shear Band Instability Under Plane Strain Conditions," *Int. J. Frac.*, Vol. 17, No. 4, pp. 389-407.

- Tvergaard, V., 1982, "On Localization in Ductile Materials Containing Spherical Voids," *Int. J. Frac.*, Vol. 18, pp. 237-252.
- Tvergaard, V. and Needleman, A., 1984, "Analysis of the Cup-Cone Fracture in a Round Tensile Bar," *Acta Metall.*, Vol. 32, No. 1, pp. 157-169.
- Walker, K.P., 1981, "Research and Development Program for Nonlinear Structural Modeling with Advanced Time-Temperature Dependent Constitutive Relationships," NASA Report CR-165533, NASA Lewis RC.
- Walters, W.P. and Zukas, J.A., 1989, Fundamentals of Shaped Charges, Wiley & Sons, New York.
- Walters, W.P., 1990, "The Shaped Charge Concept, Part 1. Introduction," Tech. Report BRL-TR-3142.
- Weihrauch, Gunter, 1987, "Armour vs. KE Rounds," MILTECH, 1/87, pp. 23-36.
- Weng, G.J., 1987, "Anisotropic Hardening in Single Crystals and the Plasticity of Polycrystals," *International Journal of Plasticity*, Vol. 3, pp. 315-339.
- Wilkins, M.L., Streit, R.D. and Reaugh, J.E., 1980, "Cumulative-Strain Damage Model of Ductile Fracture: Simulation and Prediction of Engineering Fracture Tests," UCRL-53058.
- Wilkins, M.L., 1963, "Calculation of Elastic-Plastic Flow," UCRL-7322, UC-34, TID-4500.
- Wilkins, M.L., 1964, "Calculation of Elastic-Plastic Flow," *Meth. Comp. Physics*, 3.B, Adler, Fernback and Rottenberg, Eds, ACADEMIC Press.
- Wilkins, M.L., 1970, "The Use of Artificial Viscosity in Multidimensional Fluid Dynamics Calculations," LLNL UCRL-78348.
- Wilnanski, K., 1992, "Macroscopic Theory of Evolution of Deformation Textures," *Int. J. Plasticity*, Vol. 8, pp. 959-975.
- Wiman, C.L., Meyers, M.A., and H.-R. Pak, 1990, "Observation of an Adiabatic Shear Band in AISI 4340 Steel by High-Voltage Transmission Electron Microscopy," *Metall. Trans.*, Vol. 21A, p. 707.
- Whight, T.W. and Batra, R.C., 1985, "The Initiation and Growth of Adiabatic Shear Bands," *Int. J. Plasticity*, Vol. 1, pp. 205-212.
- Yuan, W., 1990, "Response of Simulated Propellants and Explosives to Projectile Impact," Ph.D. Thesis, Mechanical Engineering, University of California at Berkeley.
- Zaloga, S.J., 1987, "Soviet Reactive Tank Armour Update," *Jane's Defence Weekly*, 23 May, p. 1011.
- Zbib, H.M. and Aifantis, E.C., 1988, On the Concept of Relative and Plastic Spins and its Implications to Large Deformation Theories. Part II: Anisotropic Hardening Plasticity," *Acta Mechanica*, Vol. 75, pp. 35-56.
- Zbib, H.M., 1992, "A Model for Elastoplastic Materials with Anisotropic and Vertex Effects," *Microstructural Characterization in Constitutive Modeling of Metals and Granular Media*, ASME MD - Vol. 32, Ed. G.Z. Voyatzis, Phoenix, AZ, April 1992, pp. 45-54.

- Zbib, H.M., 1992, "The Strain Gradient Phenomenon in Viscoplasticity: Theory and Application to Shear Instability," Proc. 13th Risø Int. Symp. on Materials Science, Eds. S.I. Andersen et al., Risø National Laboratory, Roskilde, Denmark, pp. 525-537.
- Zener, C., and Hollomon, J.H., 1944, "Effect of Strain Rate Upon Plastic Flow of Steel," J. Appl. Phys., Vol. 15, p. 22.
- Zukas, J.A., Jonas, G.H., Kimsey, K.D., Misey, J.J. and Sherrick, T.M., 1981, "Three-Dimensional Impact Simulations: Resources & Results," in Computer Analysis of Large Scale Structures, ASME AMD Vol. 49, eds. K.C. Parks and R.F. Jones, Jr., pp. 35-68.

**C. TECHNICAL METHODS FOR REDUCING
GROUND FORCES FRATRICIDE**

**William J. Dally
Massachusetts Institute of Technology
Cambridge, Massachusetts**

**Kevin K. Lehmann
Princeton University
Princeton, New Jersey**

**Robert A. Hummel
Courant Institute of Mathematical Sciences
New York University
New York, New York**

TECHNICAL METHODS FOR REDUCING GROUND FORCES FRATRICIDE

1. SUMMARY

Fratricide has historically accounted for 10-15% of all casualties [1] and was responsible for 24% of coalition casualties in Operation Desert Storm (ODS) [2,3]. The majority of the fratricide incidents in ODS have been classified as being due to misidentification or poor coordination. While most fratricide incidents involve ground forces, until recently most work on combat identification has been restricted to aircraft. A simple ground combat identification system has the potential to greatly reduce fratricide.

We have investigated technical issues relating to ground combat identification and have found that:

- An effective, low-cost ground combat identification system can be constructed today using existing commercial off-the-self technology. No new technology need be developed to field such a system.
- Near IR lasers are preferable to millimeter wave (MMW) radar for a question & answer (Q&A) IFF system. Commercial solid-state lasers, developed by the communications industry, are available at low cost and can be used to construct a system that is smaller, lighter, less expensive, and has a narrower beam than a MMW system. Power calculations show that laser IFF systems work beyond the range of an IR sight in obscured environments. An NIR interrogator capable of working to 5 km has an optical and electronic complexity similar to that of a sophisticated 35 mm camera and is simpler mechanically. A transponder is considerably simpler than the interrogator. Such systems, if produced commercially in volume, would cost less than \$500 per unit.
- A Q&A system to confirm that a target is not friendly and a position measuring, distribution, and reporting system to enhance situational awareness (SA) are both effective means for ground combat identification. A Q&A system with the interrogator integrated into a weapon's sight has the advantage of being simple to use, stealthy, self contained, and difficult to jam. Q&A interrogators can be deployed on weapons ranging from rifles to helicopters and transponders can be deployed on targets ranging from dismounted foot soldiers to tanks and APCs. A unit position information system is more complex, more difficult to use, and subject to jamming but has advantages in enhancing SA beyond reducing fratricide. Such a system can be deployed with vehicles and infantry companies. Using both systems together would provide a very reliable means for ground combat identification.
- Transponders and interrogators for a Q&A system should be tailored for particular classes of platforms and weapons while being interoperable across

units. A single unit cannot meet the weight, size, power, and range requirements for all platforms. For example, an infantry rifle must have a light, low-power, interrogator with a range of 300 m while a tank or attack helicopter can accept a heavier higher-power unit with a range of 5 km.

- Before committing to development or production of any Q&A or SA enhancing system, candidate systems should be simulated and evaluated through their use in simulated (SIMNET) or actual training exercises. Such simulations can measure the effectiveness of the systems in reducing fratricide, determine their effect on unit effectiveness, test new doctrine which takes advantage of the systems, and enhance the ergonomics of the systems.
- Non-cooperative target recognition (manual or automatic) remains a promising research area but is not yet mature enough to be used for IFF in the near term. Situations in which the type of a vehicle does not imply its alignment (friend or foe) render non-cooperative techniques suitable only for separating neutrals from combatants.

2. INTRODUCTION

2.1 The Fratricide Problem

Fratricide (a.k.a. amicide and friendly fire) has been a recognized threat since the beginning of organized armed conflict. In recent years, concern over this issue has increased. This is due at least in part to the U.S. experience in ODS, where fratricide rates were much higher (~24%) than the traditionally accepted number of ~2% of allied casualties or even a revisionist estimate that suggests that historical fratricide rates are 10-15% [1]. Several factors of ODS allowed for a more accurate assessment of fratricide than in the past. These include the near absence of enemy air power after the early hours of the conflict (making all air-to-ground attacks on coalition forces suspect) as well the U.S.-unique use of depleted uranium shells, which left an unambiguous radiation signature. Further, it is to be expected in a rout that fratricide rates will be elevated merely because the enemy is so ineffective at inflicting casualties. The percent of kills inflicted on our own, or allied, troops compared to the total number of U.S. kills is perhaps a better measure of fratricide rates, but suffers from the great uncertainty in the number of casualties inflicted on the enemy.

These reasons indicate that the fratricide rates in ODS were likely not as anomalous as they first seemed. In spite of this, there are many reasons to believe that systemic reasons exist as to why fratricide will be an increasing problem in future U.S. military engagements unless changes are made. These include:

- The increasing speed of units on the battlefield which makes it difficult to maintain up-to-date situation awareness (SA).
- The increasing range of modern weapons which give the possibility of engagement, with high probability of kill, as soon as potential targets are detected by FLIR or other enhanced detection systems.

- The increased likelihood that the U.S. engagements will be as parts of coalitions, the members of which may have essentially identical equipment as the enemy.
- The possibility is significant that the enemy will have U.S.-produced military equipment, because of changing alliances or third party transfers.

Even without an increased risk, it is morally and politically imperative that the U.S. use its advances in technology and tactics to minimize both fratricide and attacks on neutrals, while not reducing the effectiveness of our troops.

In ODS, 35 of 146 (24%) soldiers killed in action and 72 of 467 (15%) wounded in action were victims of "friendly fire." Fratricide in ODS primarily involved vehicles. Of 21 Army soldiers killed, only four were soldiers on the ground. The others were crewmen of armored vehicles. In ODS 39% of the incidents (11 of 28) were classified as due to target misidentification and 29% (8 of 28) were classified as caused by "coordination problems." A system that identifies friendly targets and provides communications information on location will go a long way toward reducing these two primary causes of fratricide in ODS.

While fratricide in ODS was primarily a vehicular problem, data show that dismounted infantry fratricide was the predominant problem in Vietnam and WWII [4,5]. A combat identification system should be designed to deal both with vehicles and with dismounted individuals.

The range at which an identification system must work is dependent on the weapon/target system involved. For armored vehicles, the system must work to a range of 3-5 km. A system for dismounted infantry need only work to a range of 300 m.

2.2 Approaches to Combat Identification (CID)

In the past doctrine and procedures, including the Rules of Engagements (ROE) have been the principle means for avoiding ground forces fratricide. Active Q&A systems have been primarily used to protect aircraft despite the fact that most victims of fratricide have been ground soldiers. In ODS, most of the victims were in vehicles, and this could be expected to continue to hold in fast moving battlefields. However, in Vietnam and WWII, most incidents of identified fratricide involved dismounted soldiers. Thus a "solution" for CID that cannot be extended to dismounted soldiers may be of limited effectiveness in jungle, mountainous terrain, or urban areas where dismounted soldiers play a larger role. Such would be the case, for example, if the United States became involved in a ground combat operation in Bosnia.

Efforts to reduce fratricide need not degrade combat effectiveness or reduce aggressiveness. A combat identification system has the potential to accelerate the identification process and to reduce the need for restrictive procedures and thus could potentially increase combat effectiveness. To design an appropriate CID system the Army

must develop appropriate supporting doctrine and ROE and evaluate the effectiveness of the overall system. For example, appropriate procedures must be developed to trade off the probability of type I (false identification of an enemy as friendly) verses type II (failure to identify a friendly) errors. Because of the disastrous effects of fratricide on morale, it has been stated that "increased fratricide is not acceptable even if total losses are reduced." [6] However, the principle that fratricide prevention should not lead to a significant increase in net casualties can hardly be considered controversial.

The issue of compliance is a major factor in the effectiveness of the system. The system is rendered completely ineffective if there is widespread disuse because our troops feel that the system reduces their capabilities or because they lack confidence in the reliability or safety of the system. In ODS, most Air Force pilots are reported to have turned off their Mark XII IFF system.

There are three major technical approaches to the CID problem: Q&A systems, positional information systems to enhance SA, and non-cooperative target recognition. A comprehensive CID system should include both a Q&A system and a positioning system to improve SA.

Question and Answer System

Question and answer systems have long been used by soldiers in close-in situations. Sentries asking an approaching soldier who won the last world series is an example. There is always the risk that friends may fail to answer (if they were not baseball fans in the above example), but such an approach is clearly better than either shooting anyone who approaches or letting enemies get close enough to open fire on an exposed sentry.

An automatic Q&A system uses an electronic interrogator associated with the shooter to query an electronic transponder associated with the target. The transponder then replies with an appropriate code if the target is friendly. The interrogation and reply are usually performed by means of electromagnetic radiation and are coded to reduce the probability of interception, jamming, and spoofing. To be effective, a Q&A IFF system must have the following properties:

- It must be reliable to earn the confidence of the soldiers.
- It must not significantly increase the signature of the shooter or target.
- It must not decrease the effectiveness (decrease range or increase firing time) of the shooter.

An automatic Q&A system could be made to both decrease fratricide and increase the effectiveness of our forces, since ROE could be changed to allow for more aggressive utilization of the range of existing weapon systems. It is imperative that any such system have the confidence of the soldier, or else it will be disengaged, as the Mark VII has been by the Air Force. Such a situation would likely be worse than no Q&A system at all.

The shooters, especially if they fear fire from the target, are unlikely to use a system that decreases the range over which they can use their weapons or increases the time required to target and fire. They also will not willingly announce their presence to other enemy units that may then direct fire on them. Obviously, firing their weapon "announces" their presence, but an electromagnetic emission can be exploited by the enemy in other ways. Even a low-probability of detection system, used in one engagement, can be discovered by "analyzing the tapes" and may likely be counterproductive in future engagements. Using very narrow beams reduces the probability of such *a posteriori* analysis.

Target vehicles do not want to increase their signature to enemy shooters. Omnidirectional beacons that announce they are "friendly" could be used by enemies as a "come kill me" signal. Attempts can be made to "hide" the beacon in the noise, but it must be remembered that it is hard to do so over a wide range, i.e., a weak signal to a distant friend will be a much larger signal to a close enemy. Also post-conflict analysis can discover the hidden signal for later exploitation. As with the interrogators, a narrow-beam transponder reduces the probability of detection and the potential for post-conflict analysis.

The Army is prepared to go ahead with an acquisition of a Q&A system based upon MMW technology. Later in this report, we propose a system based upon NIR radiation and a modulated corner reflector. In this report, we will describe the advantages of the NIR system, as well as its apparent limitations.

Positional Information System for Improved SA

A distributed information system for measuring, communicating, and displaying the positions of friendly units can greatly enhance the SA of a unit commander. Many of the components for such a system are already "in the pipe" because of command and control requirements. The GPS system, due to be universally incorporated into Army vehicles, should allow for a dramatic decrease in fratricide caused by either the shooter or the target not knowing their location, and thus a mistaken belief that no friendly should be "in the sights" of the shooter. It must be considered that a sophisticated enemy could jam either or both the GPS or the communication links that current situation awareness would require. The ideal situation would be for each friendly to broadcast their current position, in secure code, to central command and to each nearby friendly, who would make such information available to the targeting computer. The use of such a system would be extremely dangerous if it was suspected that the enemy could listen in or even spoof such a system. Further, in the "fog of war" it is unlikely that perfect situation awareness could be maintained. If gunners were allowed to shoot anything not in their database, it is likely that fratricide would increase! Effective presentation of SA data to the gunner is an important technical issue that needs to be addressed. Fighting in coalitions, such as in ODS, can carry with it a greatly increased chance of security violations which could compromise reliability on situation awareness. Also, coalition forces may not be equipped to link into such an SA system.

Even though increasing the timeliness and completeness of situation awareness through improved communication should be an important priority of the military, when a gunner lines his sights up on a potential target, it would be very desirable for there to exist a final check that would reduce attacks on friendly or neutral individuals or vehicles. While one wants a system of high reliability, it must be remembered that this situation will only come up when other methods of identification of friend from foe have been exhausted and thus fratricide should be reduced in direct proportion to percent of correct friendly identifications. This will be mitigated by changes in ROE that lead to targeting with less information than at present. As in the past, the intelligent choice of ROE requires an understanding of the realistic risk of errors, and the benefits of increased aggressiveness versus the risks of increased fratricide.

Non-Cooperative Target Recognition

Non-cooperative target recognition holds the promise of an ideal system, since it does not require the "target" to increase its "signature" in any way (which could be exploited by the enemy) and perhaps most important, it could allow for positive enemy identification. A Q&A system can only identify "friends" and would not allow enemies from neutrals to be distinguished. Despite enormous investment of resources automatic target recognition remains very limited. It is a reasonable goal to aim for systems that will aid a gunner in discriminating between a tank and a truck. Such a system could go a long way in solving the "neutral" problem since in most situations it would be reasonable that any military vehicle is either a friend or foe, not a neutral. Separation of a U.S.- vs. Soviet-built tank is a much more difficult problem. It should be remembered that when an enemy vehicle first becomes "visible" in FLIR it is only a few pixels across and possibly blurred by atmospheric distortions, and thus information content is limited. The enemy can be expected to make simple changes in its equipment to minimize any differences in "signatures" that an automatic target recognition system may be using. Finally, in the modern world it will may be the case that an M-1 tank is not necessarily friendly and a T-60 not certainly foe. Thus inclusion of ATR capabilities in our weapon systems should be explored and likely may help reduce attacks on neutrals, it is much less likely to separate friend from foe.

2.3 Report Outline

The remainder of this report addresses technical methods for ground CID. Section 3 describes a simple ground combat Q&A system based on a laser interrogator and a gated corner reflector. Adaptations of this system for both vehicular and personal use are discussed. A low-cost combat information system is described in Section 4. This system would display the locations of all friendly units to aid situational awareness. Technology alternatives and system vulnerabilities are discussed in Section 5. The report closes with some recommendations for future research in this area.

This report includes two annexes covering related substudies. One describes mesochronous coherent detection, a method for decoding a spread-spectrum interrogation signal that does not require the transmitter and receiver to be synchronized. The second annex describes SpEcBar, a frequency-domain spatially encoded marker that can be used to mark friendly units.

3. A LOW-COST GROUND COMBAT QUESTION & ANSWER SYSTEM

3.1 System Goals

The system will provide the capability to identify friendly ground targets to operators of weapons systems who are likely to commit fratricide on these targets. The system may be used both to identify vehicles (tanks, APCs, and trucks) and to identify individual soldiers. The system must have a range of at least 5km for vehicles and 300m for soldiers. It must operate in the same environmental conditions as a modern IR weapon sight. The system should not significantly increase the signature of either the target or the weapon system.

3.2 Proposed System Overview

The system consists of two components, the interrogator and the transponder. An interrogator is aligned with the boresight of a weapon. A transponder is located on the asset to be protected (a vehicle or an individual). A target is identified using an interrogate/reply protocol as shown in Figure 1. The waveforms of the interrogate and reply signals are shown in Figure 2.

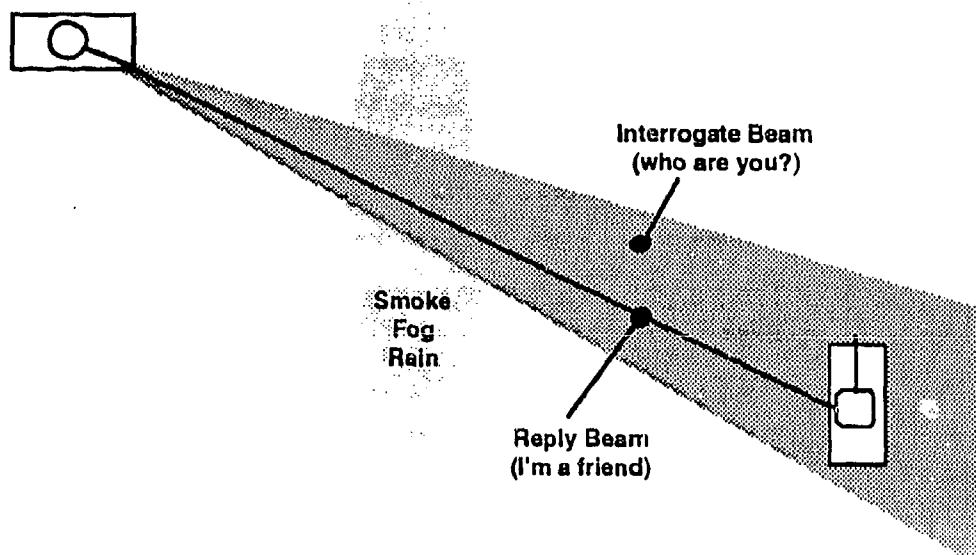


Figure 1. Overview of the Q & A System

- The interrogator transmits a narrow-beam, low-power, coded signal to the target. The signal is coded for authentication using a programmable key. The signal also includes a timestamp to prevent spoofing by retransmission and a random number to distinguish multiple queries. The narrow beam reduces both the probability of detection due to activating the CID system and the probability of misidentifying a target because several targets are in the beam pattern or its sidelobes. Also, the narrow beam signal is difficult to jam or intercept. The power levels of the system are set so that it will operate reliably even in environments obscured by smoke, fog, and rain to at least the range of the weapon/sight it is associated with.
- The transponder receives the interrogation signal, detects it, and decodes it for authentication.
- If the interrogate pulse is authenticated, the transponder opens a shutter to expose a corner reflector for a brief period of time. The delay from the interrogate pulse to the start of the exposure and the duration of the exposure are both used to code a reply. If more information is required in the reply, a ferroelectric LCD light-valve can be placed in series with the shutter to modulate the reply signal at 10 kHz.
- The interrogator illuminates the corner reflector with a low-power, narrow-beam signal and monitors its exposure. The corner reflector directs all of the intercepted energy directly back to its source so there is no additional loss due to beam spreading on the return trip. The interrogator receives the reflected signal with a heterodyne detector. This is easy to accomplish as the same laser can be used for source and reference. It only needs to be stable over a period of 30 μ s. If the interrogator sees the correct coded reply, it identifies the target as friendly.

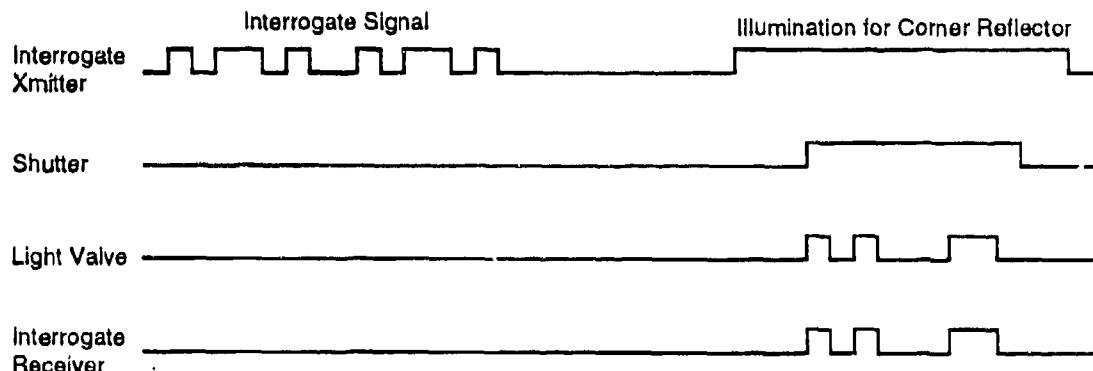


Figure 2. IFF Waveforms Showing Sequence of Operations

The interrogate signal need only be long enough to transmit the required information (see below), about 4 ms. The time for the reply is set by the minimum delay and minimum exposure time of the shutter, about 20 ms. This reply time would be increased if more reply bits are required from a slow mechanical shutter. The entire ID procedure is dominated by the reply time and thus will take about 20 ms.

3.3 Vehicular vs. Personnel Systems

Long-Range Weapons

Weapons equipped with FLIR sights (e.g., tanks, attack helicopters, and anti-tank missiles) would use an ID system that is able to operate to a slightly greater range than the lesser of the weapon range and the sight range; 5 km should be adequate for most weapons of this class. A spot size of 5 m is selected as large enough to illuminate an entire selected target, but not so large as to illuminate a nearby target. A "stepped" beam will be used to maintain a constant spot size of 5 m at ranges from 100 m to 5 km.

In adverse environmental conditions (fog or smoke) the CID system sees an attenuation factor four times that seen by the FLIR (a factor of two due to the round-trip required by the interrogation beam, and a factor of two due to wavelength). However, the power budget of the ID system is large enough that it is able to operate to a range equal to or greater than the FLIR even with the increased attenuation. The CID system has much better energetics than a laser rangefinder. Even though both require their beams to travel a round-trip, the rangefinder scatters off the target while the CID system is reflected directly back via the corner-reflector. A detailed analysis of operation in an obscured environment is given below.

The interrogator would be slaved to the sight; the laser will be centered on the target in the crosshairs. Before firing, the interrogator will ID the target. On weapons that must range the target before firing the ID can be accomplished at the same time that range is measured. The ID information can be displayed on the FLIR, displayed separately, and/or used to lock-out the weapon. To maximize effectiveness, the gunner should be able to see the ID information without looking away from the weapon's sight.

Transponders for this system would be mounted on typical targets for this class of weapon (most vehicles).

To operate effectively at the range of the FLIR, the long-range system will employ a high-power laser diode and a heterodyne detector in the interrogator. This system can also afford the extra size and expense of an LCD light valve to modulate the corner reflector at high speed.

Short-Range Weapons

Weapons with visual sights (e.g., rifles and machine guns) and limited range as well as artillery spotters can use a simpler, less expensive system. As with the long-range system, the visual ID system would operate to the lesser of the weapon's range and the sight range. For most infantry weapons, a range of 300m is adequate [5]. A spot size of 2m is adequate to entirely illuminate a target.

Interrogators would be mounted on the sight of the weapon to illuminate the target in the crosshairs. The interrogator can be manually triggered to ID the target or could be

activated by the weapon to ID the target when the trigger is pulled and lock-out the weapon if the target is friendly. As with the long-range system, the ID information should be provided in the weapon's sight so that it does not slow the operator.

Transponders for this system would be made very inexpensive and would be mounted on typical targets for this class of weapon (e.g., individual soldiers).

To reduce the cost, power, and weight of the short range system, the interrogators will take advantage of the shorter operating range by employing a lower-power laser diode, a conventional (non-coherent) detector, and a smaller aperture. The transponders will employ only a mechanical shutter (no LCD light valve) and will also have a smaller aperture. The short-range and long-range units will be interoperable in that (at the appropriate ranges) both transponder types will open their shutters in response to a query from either interrogator type and both interrogator types will respond to the appropriate shutter sequencing from either transponder.

3.4 Possible Implementation

This system could be implemented using a low cost commercial infrared laser diode (0.9μ or 1.5μ) to provide the interrogate signal. Block diagrams of the two components are shown in Figure 3 (Interrogator) and Figure 4 (Transponder).

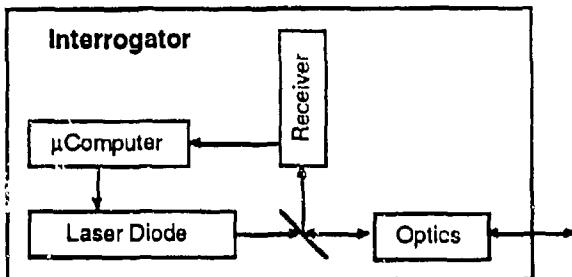


Figure 3. Interrogator Block Diagram

The laser is keyed at 10 kbits/s to transmit the codeword on top of a subcarrier. The subcarrier may be a square-wave, or a repeating codeword to hide the signal. A method for mesochronous detection of a repeating "random" codeword is discussed in the annex.

Some simple optics are required to form the laser output into a narrow beam. For a beam width of 5 m at 5 km (1 mr), a minimum aperture of 1 mm to 1.5 mm is required. The system will use a 2 cm aperture to provide adequate collection area for the interrogator receiver which shares the optical path with the transmitter. For long-range systems, the transmitter beam can be stepped by the optics so that weaker, wider-angle beams are superimposed on the main beam to keep spot size constant (at 5 m) at shorter ranges.

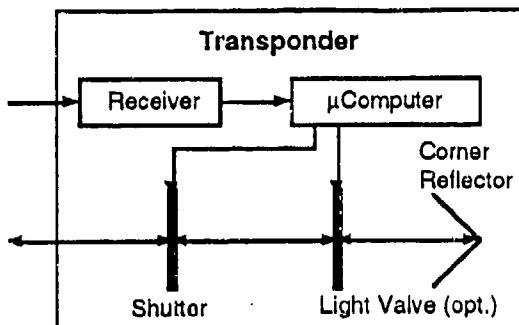


Figure 4. Transponder block diagram

The transponder receiver uses a wide-angle spherical lens to focus the incoming beam on the receiver's detector. Depending on the angle of the lens, several lenses may be required to cover the entire field of view. The receiver need not be located particularly close to the shuttered corner reflector. They both just need to be covered by the beam.

If a 0.9μ system is used, the detector can be a Si photodiode (NEP of 10^{-15} W/Hz $^{1/2}$); a thermoelectrically cooled Ge photodiode (NEP of 10^{-14} W/Hz $^{1/2}$) would be used for a 1.5μ system. An aperture of 20 mm would intercept 1.6×10^{-5} of the energy in the 5 m beam. With 5mW of beam energy, the detector would see $\sim 10^{-7}$ W on a clear day. At a bit rate of 10^4 Hz, this gives a power margin of 10^6 above detection threshold (for the Si receiver) to account for losses due to scattering in adverse environmental conditions.

Coherent detection of the electrical subcarrier is applied to the electrical output of the photodiode as described in the annex. This cancels all noise (background and jamming) that is not correlated with a codeword. In a noisy background, this will make the signal invisible to anyone not in possession of the subcarrier frequency and codeword. A single custom integrated circuit can provide both the correlation needed for coded coherent detection and the encryption/decryption necessary to authenticate the interrogate signal.

Heterodyne detection of the optical signal at the transponder would permit operation at even lower power levels (NEP of 10^{-18}) but would require very close frequency matching of the lasers in the interrogator and transponder, perhaps by using an absorption line of a gas-filled tube as a reference. This would increase expense and is not required to meet the system goals.

Heterodyne detection is very easy at the interrogator since it is sensing its own reflected signal and thus can be used there to reduce the power level of the signal during the reply period.

The transponder selectively returns the interrogation signal to the interrogator using a shuttered corner reflector. The corner reflector sends almost all of the received energy directly back to the interrogator. As with the wide-angle receiving lens, several corner reflectors may be required to cover the entire field of view. A corner reflector with a 20 mm aperture would (as with the receiver) see 1.6×10^{-5} of the beam power ($\sim 10^{-7}$ W on a

clear day). To have an effective aperture of 20 mm when exposed at an angle, the corner reflector would need to be somewhat larger.

Assuming 50% losses in the reflector and shutter and 90% losses due to diffraction beam widening, $\sim 10^{-6}$ of the beam power (4×10^{-9} W on a clear day) arrives back in the aperture of the interrogator. With a data rate of 10^4 bits/s, the heterodyne receiver detection threshold is 10^{-14} W. Thus the system has a power margin of 4×10^5 on a clear day. The power margin permits the system to tolerate round-trip scattering losses due to adverse environmental conditions of 4×10^5 a power loss of 6×10^2 in each direction. A larger interrogator aperture or corner reflector aperture would increase this margin.

The corner reflector is exposed using a fast camera shutter (uni-blitz). A mechanical shutter is adequate for sending a reply of 4-8 bits encoded in the delay and duration of the reflector exposure. A reply of a few hundred bits can be achieved (at significantly greater cost) by using a combination of a mechanical shutter (to get very good contrast in the off state) and an LCD light valve (for fast signaling at lower contrast during the reply). The combination would be used by opening the mechanical shutter for a fixed exposure time (perhaps 10ms) and modulating the reply signal during this interval using the LCD light valve.

The interrogator receiver must be located in the transmitter optical path to intercept the returned beam. This can be accomplished with a polarized beam splitter. The receiver can share the transmitter's beam forming optics. The interrogator receiver can easily employ a heterodyne detector (since it is detecting its own signal) by locking the transmit laser to the receiver local oscillator.

The shutter and the other system components are controlled by a single chip microcomputer. The interrogator microcomputer would upon user command generate the interrogation signal, perform automatic setting of power level, monitor the reply signal, and perform encryption/decryption. Depending on the code chosen, a separate semicustom integrated circuit may be required for the encryption/decryption.

The transponder microcomputer would monitor the receiver, decrypt and authenticate potential interrogation signals, and (if the interrogate is authenticated) control the shutter(s). A semicustom integrated circuit would be used to perform the mesochronous coherent detection of the interrogate signal and the encryption and decryption functions. The code for each microcomputer should be quite small (a few kbytes) and fit entirely on an on-chip ROM.

This system can be built inexpensively using a standard communication or CD laser diodes, standard photodiodes, a simple corner reflector, a standard camera shutter (and possibly a standard LCD light valve), a standard microcomputer, and (if required) a semicustom integrated circuit.

To keep the interrogator from increasing its signature, power level can be increased exponentially from a low level until either a reply is detected or some threshold is reached. The threshold can be a programmable function of range and environment. The interrogate signal should be kept short to minimize exposure when identifying unfriendly targets. To keep the transponder signature low, a very low maximum duty factor on shutter open time should be set.

3.5 Signal Content

The interrogate signal should include the following information in encrypted form:

- A timestamp with 50 ms resolution and a scale of at least a day (20 bits would suffice). To validate the timestamp requires that the timebases of all interrogators and receivers be synchronized to at least 50 ms accuracy. Alternatively, a scheme using relative time could be employed requiring looser synchronization of timebases. A low-cost, uncompensated crystal oscillator (accuracy typically 1 ppm) would be sufficient to keep this time base if reset to a master clock once per day.
- A random number to make the reply dependent upon information unavailable to any adversary (16 bits). This can be generated with a noise diode or a pseudo-random number generator. The random number also allows two simultaneous interrogations to be distinguished.
- Optionally, other information: the interrogator's ID, position, status, or other communication information can be included as required.

Overall, the interrogate signal will require at least 36 bits. These bits should be encrypted as a block to prevent an adversary from inserting or extracting any of the parts. A key of sufficient length to prevent systematic attack (64 bits) should be used for this encryption. The key should be changed daily. This can be done automatically at the same time the interrogators and transponders have their timebases set. A handheld service unit with a wireless IR interface can be used to program and validate the timebase and encryption key, or this programming can be performed via a standard communications link. In either case the programming and validation can be performed without removing the unit from its context.

The microcomputer itself can generate the interrogate signal. A semicustom integrated circuit may be required for encryption. If the interrogator supports a bit rate of 10 kbits/s, the interrogate signal can be transmitted in under 4ms. It should be framed to allow the receiver to identify the start of the signal.

The transponder decrypts all interrogate signals received from its receiver and validates both the codeword and timestamp fields of the signal. If these signals are validated, the transponder generates a reply containing the following information:

- The random number included in the interrogate signal. This allows the interrogator to assure that the reply is to its signal (16 bits). If a mechanical shutter is used, a 4-8 bit signature of the random number can be sent.

- Other information: transponder ID, position, status, and other communication information.

A timestamp is not required on the reply signal since it must be sent a fixed delay after the interrogate signal. Encryption is not required on the reply signal since the receiver decryption guarantees that an adversary cannot derive the random number from an adversary pulse. A low cost system with a simple mechanical shutter can encode 2-4 bits in each of the delay and duration (4 bits assumes we can control event times with less than 2 ms jitter, 2 bits assumes 8ms jitter).

3.6 Atmospheric Attenuation

Consider an target object of cross sectional area A , a distance D from a tank. If the target is at ΔT above the ambient background temperature T_0 , then the total IR power emitted from the target is on the order of

$$P = 4 \sigma_b T_0^3 \Delta T A$$

where $\sigma_b = 5.5 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$ is the Stefan-Boltzmann constant. The fraction of this radiation that will be intercepted by the FLIR sight with aperture a will be $\sim t (a/2D)^2$ where t is the atmospheric transmission. The HgCdTe detectors in the FLIR will detect a certain fraction f of this black-body radiation that will be in the detection bandpass, leading to an expression for the expected detectable power of:

$$P(\text{det}) = \sigma_b T_0^3 \Delta T A t (a/D)^2 f$$

If we assume values of $T=300^\circ \text{ K}$, $\Delta T=10^\circ \text{ K}$, $A = 5 \text{ m}^2$, $t=1$, $a=5 \text{ cm}$, $D= 5 \text{ km}$, $f=0.5$, we get an estimate $P(\text{det}) = 4 \times 10^{-9} \text{ W}$. This is close to the noise equivalent power of a cooled HgCdTe, so this is consistent with an estimated range of 5 km for the FLIR sight in favorable weather conditions. At this range, the NIR Q&A system is expected to have a Signal to Noise (S/N) of $\sim 10^5$ under good visibility.

In order to consider the effect of atmospheric, lets look at the effects of heavy rain (16 mm/hr). The loss to mid-IR near $10.6 \mu\text{m}$ is expected to be $\sim 9 \text{ dB/km}$. This will reduce the range of the FLIR sight from 5 to about 1.3 km. The attenuation of the NIR will be slightly less, about 8 dB/km . At the range of 1.3 km, the expected S/N for the Q&A system will still be $\sim 10^4$ despite the two way attenuation. Under the same rain conditions, the attenuation for 95 GHz is about 7.4 dB/km , very similar to that of the NIR, thus rain does not give any significant advantage to the millimeter wave system.

In heavy fog, the attenuation of the mid-IR is about 30 dB/km . Under these conditions, the range for the FLIR sight will be reduced to about 600 m. The attenuation for the NIR will be about twice that of the mid-IR, so about 60 dB/km . As a result, at 600 m the two-way attenuation will be about 10^{-7} , but the decrease in diffraction for the range of 600 m verse 5km, means the S/N for the Q&A system will be about unity even in

heavy fog. Under these same conditions, the attenuation for 95 GHz is only about 0.35 dB/km, and thus an MMW system would still be usable beyond the range of the FLIR.

Other important sources of atmospheric attenuation on the battlefield are blowing sand and smoke. For sand, the typical particle size will be larger than 10 μm , and thus the relative attenuation of mid-IR and NIR will be similar, like rain, thus assuring that the range of the Q&A system will be greatly exceed that of the FLIR sight. We expect a similar behavior for smoke.

In conclusion, the NIR Q&A system will have a range that exceeds that of the FLIR sight in both clear conditions and even in the presence of heavy rain. In rain, the atmospheric attenuation is roughly the same from 95 GHz through the visible. Under conditions of heavy fog, the range of the NIR Q&A system will still at least match the range of the FLIR, despite the higher loss coefficient and double atmospheric path length. Thus the NIR Q&A system should not reduce the range of engagement of the weapon system on which it is mounted. It is expected that the system will still be able to function at least as far as the FLIR sight in the presence of smoke and/or dust as well.

4. A LOW-COST GROUND COMBAT POSITION INFORMATION SYSTEM

4.1 System Overview

A simple position information system would construct and distributed a distributed database containing:

- Positions and identities of friendly units in the region of interest
- Suspected positions and identities of enemy and neutral units that have been spotted
- Terrain and environmental features.

Such a system has many uses besides fratricide reduction. However, this discussion is limited to the use of a position information system for CID.

When engaging a potential target, the shooter would place his weapon's sight on the target and range the target (e.g., using a laser range finder). A microcomputer uses the shooter's present position as well as the bearing and range information from the sighting to compute the location of the target. It then queries a local copy of the database to determine if any known objects are in the target area. A display is provided in the weapon's sight relaying (1) whether the target is friend, enemy, neutral, or unknown and (2) any additional information on the target that has been recorded in the database.

If incomplete information exists on the target, the shooter can enter additional information to be added to the target record. To keep the database up to date, each equipped friendly unit periodically broadcasts its position and velocity over a network to all other units. At the same time it also broadcasts any new or updated information on neutral or enemy targets.

In addition to providing alternative means for identifying targets in a weapons sight, a position information system would also automate position reporting for artillery spotters and thus reduce coordination errors which could result in fratricide due to indirect fire.

To aid general situational awareness and coordination, the system would generate a moving-map display of the area around the unit as depicted in Figure 5. The map would indicate local topographical features as well as positions of friendly, unfriendly, and unknown units.

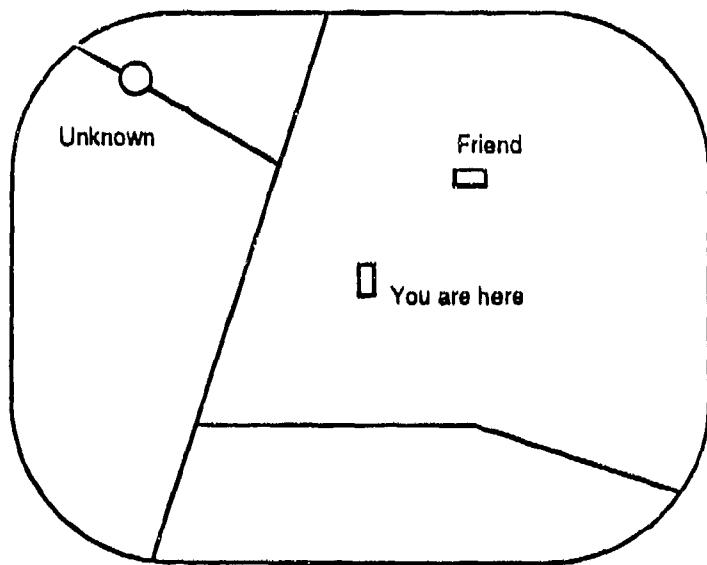


Figure 5. Moving-Map Display From a Position Information System

4.2 System Components

A position information system consists of three main components as shown in Figure 6: a positioning system, a communication system, and a database and display computer. As suggested by the figure, off-the-shelf GPS receivers, laptop computers, and cellular telephones could be used for these purposes. The information system must also be interfaced to weapon sights and spotting scopes to be used for target identification.

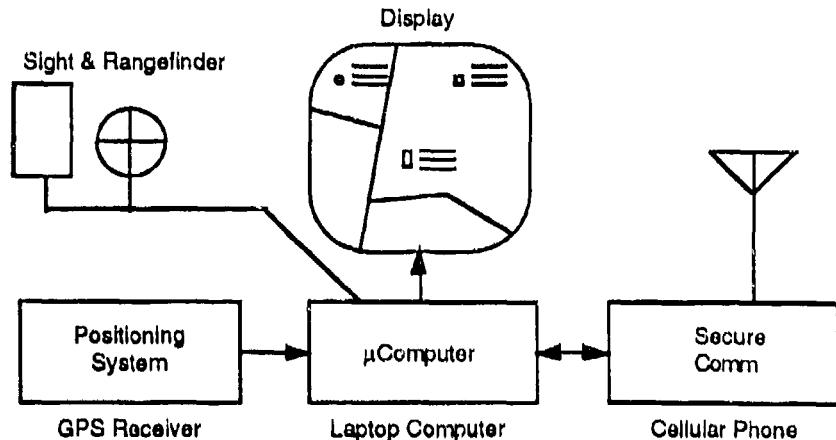


Figure 6. Components of a Position Information System

Positioning System

The positioning system must be capable of measuring the position of each vehicle to an accuracy of a few 10s of meters. A GPS receiver is clearly adequate for this task and is available commercially at very low cost (\$500). GPS, however, is subject to local area jamming. An inertial navigation system could supplement GPS and provide the capability to "ride out" GPS outages but is rather costly. Another alternative, also subject to jamming, is a terrestrial radio triangulation system.

Communication System

Once each unit has measured its position, it broadcasts this information, along with any information it may have gathered on other targets or on battlefield conditions to all units in the local area. To reduce the amount of data to be sent, positions can be sent relative to a "battlefield center point." If positions are reported on a 10 km square grid surrounding the center point to 1m accuracy, 28 bits are required to report a 2-D position. In addition, 28 bits of velocity information must be sent along with 16 bits of unit identification for a total of 72 bits. Rapidly moving units should rebroadcast their positions once per second. Velocity information is used to extrapolate positions until the next update. If there are 1,000 units on the battlefield, a data rate of 72,000 bits per second is required, well within the capabilities of a wideband UHF radio link. The required data rate can be reduced by several orders of magnitude by having units report their position only when it deviates from a linear predictive model of their position (an extrapolation given their last position, velocity, and acceleration) by more than a predefined threshold (e.g., 10 m).

Any position information system must be encrypted to prevent an adversary from gaining information as to friendly units' positions or injecting false information into the network. This encryption can be performed by the database and display computer, perhaps with the addition of a small amount of hardware. Care must also be taken in the design of the system to prevent the radio transmissions needed to broadcast position updates from

increasing the signatures of friendly units. Also, these broadcasts are subject to jamming which could disrupt position updates.

Summary information on positions reported on a single 10 km square battlefield can be transmitted up a hierarchy of position information systems to provide information on unit positions over a wide area and across boundaries as illustrated in Figure 7. For fratricide reduction purposes, only the local information is of interest.

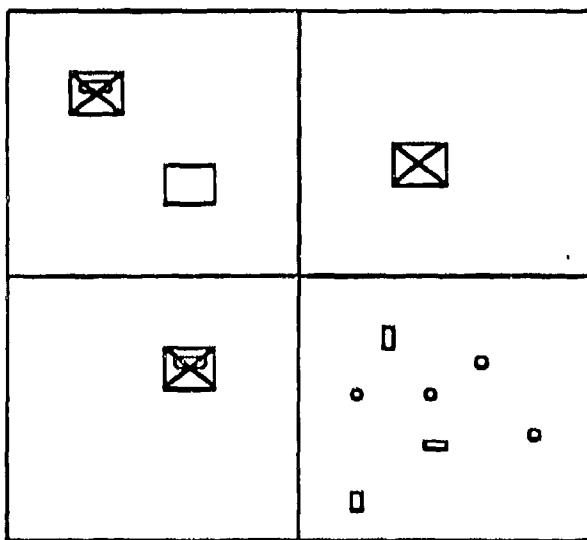


Figure 7. Detailed Information Is Directly Reported on Units in the Local Area (lower right) While Summary Information Is Relayed Indirectly for Units in Adjacent Areas

Database and Display Computer

The database and display computer receives the broadcast position updates, stores them in a local database, and displays the situation on a moving map display. The computing and storage requirements of this unit are easily met by a commercially available laptop personal computer (<\$2,000). To store 1,000 units positions and identities requires less than 20 kbytes. A standard 640x480 unit LCD display has sufficient resolution to display symbology showing terrain features and the positions of all known friendly, enemy, or neutral units. Local topographic information can be encoded into a few tens of kbytes. A database of topographical maps can be stored locally on disk or CD ROM or can be broadcast via a communications link.

System Integration

To be useful for target identification and spotting, the database and display computer must be interfaced to the weapon's sight or spotting scope. The computer must receive bearing and range information from the sight to locate the target relative to the unit's position. Also the computer should be able to display its result in the sight so that the weapon's operator does not have to look away from the sight.

4.3 Integrating a Position Information System with a Q&A System

An effective IFF system can both be assisted by, and can contribute to, situational awareness. In an ideal situational awareness system, a database is maintained and is accessible by all friendly forces, providing identification and position information of all objects, stationary and nonstationary, and especially an indication of whether those objects are threats or assets. Further, ideally, the information that is provided to an inquiry gives relative position information.

Situation Awareness for IFF

Suppose that an effective situational awareness system is available. An IFF system is then, in part, obviated, since a potential shot can be arrested by accessing the database, to determine if the intended target is friendly, a known neutral, a known unfriendly, or unknown. Different operating parameters can lead to graded doctrines, which could be preprogrammed in advance of the conflict, reflecting the environment and conditions of the engagement.

On the other hand, an IFF system may still be warranted, even if an SA system is deployed. First, the SA system may be less trustworthy than a direct signal from a sited object. In essence, a friendly about to be fired on might prefer to communicate directly with the targeting system as opposed to updating a database of positions, which the shooter is supposed to access. Accordingly, an IFF system provides an important backup capability, necessary to any practical SA system. It is especially important to have backup capability because no SA system is going to be 100% current: staleness of the database can certainly contribute to possible fratricide. By contrast, the IFF system is subject to far fewer possible confusion effects, due to its immediacy. Second, the IFF system may be a more secure communication path, in a situation where communicating with a stored database might involve less direct transfer of information. Last, because of its local and specific nature, an IFF system is theoretically faster than an SA system. This is true either because the communication is typically shorter, or even for an SA system where the database is directly accessible locally, because the size of the database that is being accessed requires a search for information that takes a certain minimum amount of time.

IFF for Situational Awareness

It is assumed that a network of friendlies are contributing to a grid of information that maintains the database. It is irrelevant as to whether the database is distributed, centralized, maintained with respect to global coordinates, or redundantly encoded using differential (or relative) coordinates. It does matter, however, that the database requires constant updating, and that not all links are necessarily reliable. Thus, every piece of information should be exploited and incorporated into a state model of the battlefield.

Each IFF query-and-answer between mutual friendlies involves a communication. There is a greater bandwidth in the query, and so the interrogator can easily transmit

coordinate information and unit identification to the interrogated platform, if it so wishes. Even in the absence of this additional data, there is usable information in the query/answer system.

Since the interrogator knows its own position (either using GPS or local inertial models), and the direction of query is known, the location of an unidentified object is constrained to a locus of locations lying on a line. Approximate range information may also be available. If two or more non-collinear friendlies perform a query on the same object, then an approximate location of that object can be formed. Of course, errors are possible if the same objects are not queried, or if there are timing discrepancies. However, three or more queries to the same object in the same time period greatly reduces the potential for error.

An answer from the queried system establishes the identity of the queried object and can provide increased accuracy of the queried platform. The queried system receives some information, but in the initial round, it is unlikely that the IFF system will reveal exact coordinate location, even though the query is encoded. Basically, a friendly that is painted by a mutual friendly "knows" that another friendly is around and may have general direction information, but will not likely be able to deduce accurate position information from the initial communication.

However, once two friendlies have established contact through an initial "handshake," they can then exchange more information in order to verify their finding and update the SA system. This could work as follows:

- The interrogator recognizing from the first round that a friendly has been targeted, composes a message to the targeted system that contains the following information:
 - I confirm that you are a friendly
 - My current position is (x,y,z)
 - My identity is ID
 - Respond (or don't respond) with your position.
- The responding system, after breathing a sigh of relief, can respond with its relative position (relative positioning requires fewer bits of information) if so requested. Relative position can be computed by finding the difference between the broadcast position of the interrogator and the known current location of the responder, obtained from GPS or any other system. Further, the responding system might in this round give more identity information, if this was not part of the first round. In any event, a minimal confirmation pulse will be sent. Note, however, that the return communication information requires a very few number of bits.

After the second communication cycle, each of the friendly parties knows more about the identity and relative (or absolute) position of the other party. This information can be fed by either party into the SA grid. There is little or no benefit to independent

communication with the grid, since the handshake protocol establishes a common information base between the two. However, independent communication with the grid is only harmful if the bandwidth of the SA system is an issue.

The SA system does not necessarily have to believe all the information fed to it. It can use Kalman filtering of the state space of the system in order to update location and identity information, as it is received, of the time-varying location/velocity/identity of the situation awareness.

5. TECHNOLOGY TRADEOFFS AND EVALUATION

5.1 Q & A System Alternatives

Our proposed NIR Q&A system uses a NIR active narrow-beam interrogator with a gated passive corner reflector. Several alternative configurations are possible [6,7]:

Active Reply Transponders

An active transponder (either laser or radar) has a power advantage in obscured environments because the beam energy undergoes one-way rather than two-way attenuation. However, active reply transponders have a directionality problem that more than negates this advantage. It would be extremely difficult to detect the direction from which the 1mr interrogation beam is coming from and to steer a narrow beam back in the same direction with 1mr accuracy in the millisecond timescale required by an IFF system. Thus, an active system would need to rely on a wide-angle, possibly omnidirectional, beam. Wide beams present a power problem and a detection problem.

If an omnidirectional beam is used, only a fraction $f = (a/d)^2/4\pi$ of the beam energy is collected by an aperture of size a at distance d . For a 2cm aperture at 5 km, only $f=10^{-12}$ of the energy is collected. Compare this with the 10^3 additional attenuation of round-trip vs. one-way transmission in heavy fog.

A wide or omnidirectional beam is much more likely to be detected or intercepted than a narrow beam because of the large amount of power being indiscriminantly radiated in all directions. Thus it presents a greater danger of increasing the signature of the target. It is also more vulnerable to being intercepted and collected on tape for later analysis. In contrast, a narrow reply beam is only visible along the line of sight to the interrogator and thus is much less likely to be detected or intercepted.

It has been suggested that omnidirectional transponder replies and beacon signals can be "hidden in the noise" using spread spectrum and other low-probability of intercept (LPI) techniques. The problem with this approach is that what is hidden in the noise at one range is well above the noise floor at a closer range. Consider an omnidirectional signal with an SNR of 10^{-2} at 5 km with a 2 cm aperture. To an adversary at 500 m, this signal has an SNR of 1 and to an adversary at 50 m, it has an SNR of 10^2 . If one is going to use LPI techniques, it is better to do so with a narrow beam in any case.

Beacons

Beacon systems, while they eliminate the need for the shooter to radiate any energy, make the target even more vulnerable to detection by an adversary than omnidirectional transponders. Not only is the beam radiated in all directions, it is also radiated all the time. Compared to the transponder, the signature is increased by an amount proportional to the inverse of the transponder duty factor.

Ungated Reflectors

An ungated reflector also increases the signature of the target by an unacceptable margin. An adversary could easily detect these reflectors by illuminating his field of view with a source co-located with his detector. The gated corner reflector system is much less observable because of its very low duty factor.

Radar vs. Lasers

MMW radar has lower attenuation than laser in smoke and fog. However a low-power (10 mW) NIR laser has adequate performance (range greater than that of a FLIR) in obscured environments and has many significant advantages.

The higher wavelength permits a tightly collimated (1 mr) beam to be produced using a very small aperture (1 mm) and this beam has no appreciable side lobes. A laser beam can be easily stepped using simple optics to provide a constant cross-section of detectable power over a wide range of distances. The result is a system that has a lower signature and permits more selective illumination of potential targets. A laser system with a 1 mm aperture can selectively illuminate one of two tanks separated by 20 m at 5 km. A 3 m antenna would be required to accomplish this with 95 GHz radar.

A laser system using commercial laser diodes and detectors is less expensive than a comparable radar system, is more compact, lighter, and requires less power. The power, weight, and size advantages are particularly important in protecting dismounted infantry where the unit must be carried by a soldier.

Modulated Directional Reflectors

Phase-conjugate mirrors have the potential to direct a narrow beam directly back to its source with a power gain of as much as 10^3 . Unfortunately the only mature technology for phase-conjugate mirrors (barium titanate) is far too slow for this application (minutes vs. ms).

5.2 Vulnerabilities

Jamming

An adversary could try to defeat the system by jamming the battlefield with a high-powered IR laser. Jamming a transponder's receiver is difficult because the adversary is sending an uncoded signal that will be largely rejected by the receiver's subcarrier coherent

detection. However, adequate headroom in the receiver must be provided to perform this rejection. Also, the adversary's wide CW beam must compete with the interrogator's narrow, pulsed beam making the energetics unfavorable. Finally, any adversary performing such jamming is easily detected and targeted.

Jamming the interrogator receiver is more difficult than jamming the transponder because the heterodyne detection will reject all optical signals that are even slightly off the frequency of the transmitted signal. An adversary could, however, litter the battlefield with corner reflectors. Nearby reflectors could then swamp a distant target. This could be countered by gating the received signal for range or collimating the beam more tightly in the vertical axis to avoid reflections from nearby ground reflectors. Using an adjustable spot beam rather than a stepped beam would also reduce vulnerability to this type of attack.

Spoofing

An adversary cannot easily spoof the system without having the cryptographic key. The use of a timestamp in the query makes every query valid exactly once. (By the time the same timestamp comes up the next day, the key has changed). The return of the query's random number in the reply makes it unlikely that an adversary can guess the correct reply without having decoded the query.

Captured Units

To prevent an adversary from removing the transponder or interrogator from a captured or destroyed unit, the system should be designed so that removing a box from the unit to which it is associated erases the key. Care must be taken in the design of this interlock to make it difficult to inadvertently erase the key. A unit should signal that it is disabled so a soldier who inadvertently erases a key knows not to trust the unit.

Exploitation

The system does momentarily increase the emissions of the shooter and the reflectivity of the target. An adversary could exploit the latter by constantly illuminating the scene with an IR laser. The adversary could then detect the momentary (20 ms) flash when a target uncovers its corner reflector. As in the case of jamming above, any adversary employing this technique is easily detected and targeted. Exploiting interrogator signals is difficult as it requires the adversary to be along the line of sight.

Multiple Simultane Queries

The system must respond properly to several near-simultaneous queries. The low duty factor of the interrogators reduces the probability of simultaneous queries. An interrogation signal lasts 4 ms and is most likely performed less than once per second. As a further measure, nearby units could be assigned 4 ms time slots during which they are permitted to send their interrogations. Waiting up to 40 ms to send the interrogation would not significantly add to the delay of the identification process.

If queries are spaced close together but do not overlap, the transponder can detect them, but may not have time to respond to both. In this case, the transponder could send a special code (incorporating the current time) that would be recognized by all of the interrogators.

6. CONCLUSION AND RECOMMENDATIONS

This report has suggested a two-part technical approach to reducing ground-forces fratricide problem: a laser-based Q&A system augmented by a position information system.

An NIR-laser-based Q&A system provides a simple, low-cost, standalone ground combat identification system. Commercial laser diodes provide a low-cost high-quality signal source. The system should employ a narrow, stepped-beam interrogator for power-efficiency and to reduce the interrogator signature. A transponder using a gated corner reflector provides a very narrow reply beam that reduces the signature of the target. The directionality of this beam more than compensates for the power required for the round-trip transmission associated with a passive transponder. The system should use an encrypted signal that includes a timestamp to guard against spoofing. The procedures used to program the key and synchronize the timebase should be simple and fail-safe so that soldiers will develop enough confidence in the system to use it.

An NIR-based transponder and interrogator have electrical and optical complexities comparable to those of an automatic 35 mm camera. If produced commercially in quantity, they should have costs comparable to those of a camera.

To be effective, different interoperable versions of the system should be deployed on different classes of weapons and targets. A system with a range of 5 km is required for long-range weapons with FLIR sights while a smaller system with a range of 300 m is adequate for many infantry weapons. Small, short-range transponders can be deployed on individual soldiers.

A laser-based Q&A system can be complemented by a position information system (PIS) to improve general situational awareness. A PIS would provide combat identification in situations such as indirect fire where a visual line of sight on the target is not available. It would also serve as a cross-check on the Q&A system.

A PIS consists of a positioning system, a database and display computer, and a communications network. The positioning system provides each unit with its own position. This position is included in a database of unit positions and geographical information maintained by the computer. This information is presented to the user as a moving map display with target identification information additionally superimposed on a weapon sight. The communication network serves to distribute information on friendly and unfriendly unit positions and other intelligence information.

A simple PIS could be constructed from a GPS receiver, a laptop computer, and a cellular phone. The computer could provide the encryption required for security. Produced commercially, such a system would cost a few thousand dollars.

We have suggested particular technical methods for reducing ground-forces fratricide. Further, these methods require no new technology development. The Q&A system uses commercial off-the-shelf laser diodes and detectors while the PIS system can be built using available GP₁ systems, laptop computers, and communication devices. However, it is premature to start large-scale manufacturing or deployment of such systems. A number of questions of effectiveness and methods of use should be answered via experimental study before such a deployment. In particular, we recommend the following actions before full-scale development:

- An NIR laser and heterodyne detector should be tested under a variety of low-visibility conditions (fog, dust, smoke, etc.) to validate our theoretical power calculations and quantify the attenuation seen under these conditions. A full Q&A system need not be developed for these tests. All that is needed is the laser, some simple optics, a reflector, and the detector.
- Models for Q&A system and the PIS should be added to simulations to determine their effectiveness. A SimNet tank battle using simulated Q&A systems could assess how soldiers react to the system and the effectiveness of the system in reducing fratricide under various simulated failure rates. These simulations could also be used to experiment with doctrine that takes advantage of the improved situational awareness provided by the two systems. We would expect that significant changes in system requirements would come about as a result of these simulations.
- A more realistic simulation of the system should be tested on an instrumented test range, such as the NTC, to test its effectiveness under more realistic conditions.

A cryptographically protected Q&A system such as the one suggested above need not be kept secret. The specifications could be published openly and commercial manufacturers could compete in the open marketplace to produce low-cost units. Such units could be safely exported as without the current key, possession of a unit would not enable an adversary to spoof or exploit our system. There are even methods whereby we could exploit the use of such "export units" by an adversary.

REFERENCES

1. Sa'adah, Col. David M., *Friendly Fire, Will We Get it Right This Time*, Office of the Surgeon General, Department of the Army.
2. Department of Defense, *Conduct of the Persian Gulf War*, Final Report to the Congress, April 1992.
3. Forecast International, *Conduct and Lessons of the Persian Gulf War*.
4. Beyer, James C. ed., *Wound Ballistics*, Office of the Surgeon General, Department of the Army.
5. Joint Technical Coordinating Group for Munitions Effectiveness, *Evaluation of Wound Data and Munitions Effectiveness in Vietnam*, Aberdeen Proving Ground.
6. Army Science Board, *Summer Study on Ground Combat Identification*.
7. Department of the Army, *Operational Requirements Document for the Battlefield Combat Identification System*, October 1992.
8. U.S. Congress, Office of Technology Assessment, *Who Goes There: Friend or Foe?*, OTA-ISC-537, (Washington, DC, U.S. Government Printing Office, June 1993).

Annex A
CODED MESOCHRONOUS COHERENT DETECTION

**William J. Dally
Massachusetts Institute of Technology
Cambridge, Massachusetts**

Annex A

CODED MESOCHRONOUS COHERENT DETECTION

Coded mesochronous coherent detection is a variant of coherent detection in which transmitters and receivers use the same frequency but have an arbitrary phase difference. With mesochronous detection there is no need for a central timebase and no need to synchronize the transmitter and receiver. The modulating waveform is also coded to distribute its energy over the frequency spectrum.

Coded mesochronous coherent detection is particularly well suited for IFF Q&A systems. In these applications, there is a need to distribute the signal energy across the spectrum to avoid detection. However, with only a short query and reply, there is insufficient time to synchronize the transmitter and receiver using conventional techniques such as a sliding frame.

Digital coherent detection involves multiplying a slowly-varying digital signal, x , by a periodic zero-mean digital waveform, w (typically a unit square wave where high is interpreted as +1 and low as -1). The modulated signal, $y = xw$, is then transmitted through a noisy channel resulting in a received signal, $y' = xw + n$, where n is additive noise. This signal is then multiplied by the reciprocal of the zero-mean waveform, $w' = w^{-1} = w$, to give $x' = x + nw'$. The received signal, x' , is integrated over the bit period, t , to give $x'' = xt + n'$, where

$$n' = \int_0^t nw' dt.$$

This signal flow is shown in Figure 1.

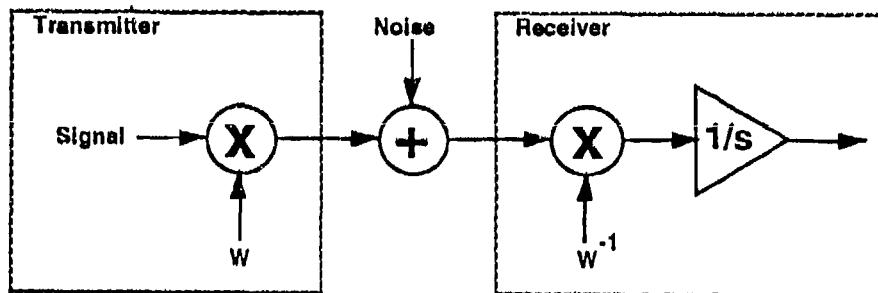


Figure 1. Signal-Flow Graph for a Coherent Detector

If n is uncorrelated with w' , n' is a zero-mean random variable with variance that increases as $t^{1/2}$. Thus by integrating for a sufficiently long period of time, the signal-to-noise ratio (SNR) can be increased by an arbitrary amount.

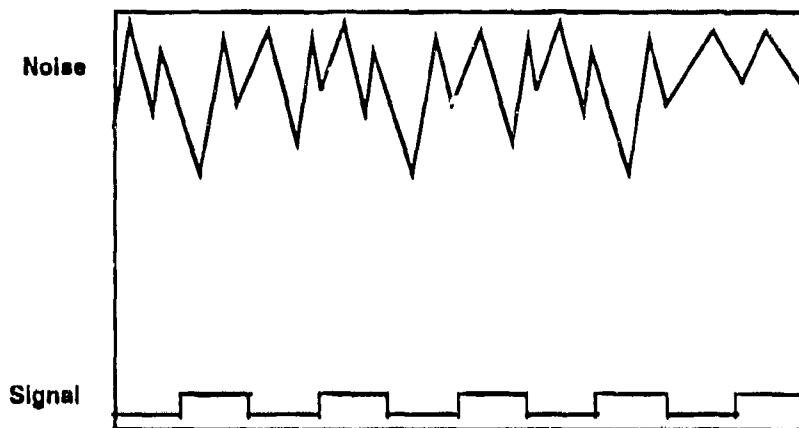


Figure 2. Mesochronous Coherent Detection Permits a Signal To Be "Hidden" Beneath a Noise Floor by an Amount Proportional to the Square Root of the Integration Time

Coherent detection can be used for IFF by transmitting the signal y at a sufficiently low power level that it is undetectable when mixed with background noise at a receiver, $|y| \ll |n|$ (Figure 2). If the SNR ($|y|/|n|$), is kept below detector thresholds, signals can be exchanged between friendly units with minimal probability of detection by the opposition (assuming the opposition does not know w). By integrating for a sufficiently long period of time, t , the integrated signal, x_t , can be made sufficiently larger than the integrated noise, n , to be detected reliably.

To spread the energy in the transmit and reply beams over the spectrum and to defeat the use of a captured unit by the opposition, a pseudo-random sequence is used for the digital waveform, w . This sequence can be changed periodically for security.

To operate the IFF units mesochronously a digital signal of length L bits (each bit representing +1 or -1) is used for the waveform, w . To receive this signal coherently, each receiver simultaneously accumulates $4L$ sums of wy' ($w^{-1} = w$ in this case). One sum is computed for each quarter-bit phase shift possible across the entire length of w . With this approach, illustrated in Figure 3, one of the sums is guaranteed to be within a quarter bit of being in phase and hence receive at least half the power of a fully coherent signal.

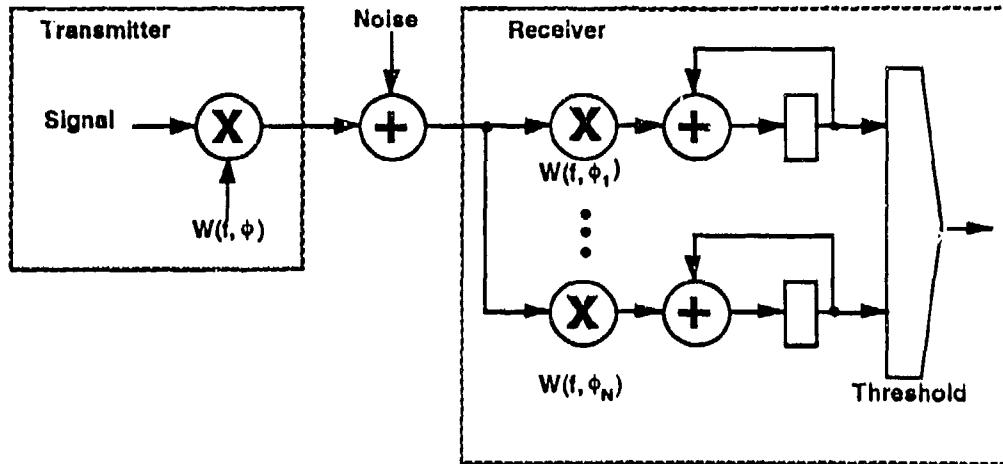


Figure 3. A Mesochronous Coherent Detector Simultaneously Detects the Received Signal with all Possible Phases in Quarter-Bit Increments and Thresholds To Detect the Proper Phase

Consider the case where the sequence length, L , is 2048 and the subcarrier bit rate is 1 MHz. The mesochronous detector must compute 10^4 sums each microsecond for a rate of 10^{10} adds per second. This is easily achieved with a small, inexpensive CMOS integrated circuit. A 16-bit adder capable of operating well in excess of 10 MHz can be constructed in an area of about 0.05 mm^2 in an $0.8 \mu\text{m}$ CMOS process. Performing the required sums requires 10^3 of these adders for a total arithmetic area of 50 mm^2 . With additional support circuitry a mesochronous detector capable of operating on sequences of up to 2048 bits at data rates of up to 1 MHz will easily fit on an inexpensive $8\text{mm} \times 8\text{mm}$ CMOS chip. For comparison, a typical high-end microprocessor chip is over 150 mm^2 in area.

Annex B
IFF WITH THE SPATIALLY ENCODED BAR CODE (SpEcBar)

Robert A. Hummel
Courant Institute of Mathematical Sciences
New York University
New York, New York

Annex B

IFF WITH THE SPATIALLY ENCODED BAR CODE (SpEcBar)

This annex provides an alternative IFF design, which like the approach described in the paper is low-cost, simple, and functional. The approach described below, which we call the SpEcBar, is simpler yet, and less sophisticated. The approach is a "marker" and adds a decal to the friendly vehicles and assets. The marker is hidden by its encoding. However, if an adversary learns of the use of the marker, they could likely quickly interpret the code, and in any case learn how to look for markers, rendering the method counter-productive. Accordingly, the SpEcBar is very likely an idea that would best be used one-time only in a single conflict.

1. IFF WITHOUT TRANSPONDING

Many of the proposed IFF schemes are based on modifying the signature of the friendlies in such a way that the enemy is unable to see the change, but that friendlies, who know where to look, can easily see the modification. One such scheme was employed in ODS, involving blinking IR radiators. The ODS scheme was crude and depended on the fact that the adversary lacked night vision. More generally, proposed schemes wish to hide the signature modification in background noise, so that the pattern is detectable when sought, but is not readily observable by ordinary sensors. Generally, such schemes will be useful when kept secret, but might require modification after discovery. If the scheme is low-cost, it might provide reliable IFF for one conflict.

The signature modification method results in an IFF system without a need for transponding. This offers the advantage of simplicity, and greater reliability. The danger is that the scheme will be discovered, and that the modified signature will give a good clue for the adversary. Active, coded response is probably preferable to a "Made in the USA" placard. Nonetheless, there will be occasions where a simple, non-transponder marker-based IFF system will have utility. We next describe a system, called the SpEcBar, which provides an extremely simple, low-cost, signature modification method for identifying friendlies.

2. THE SPATIALLY-ENCODED BAR CODE

The Spatially-Encoded Bar Code (SpEcBar) uses distributed basis functions, such as sinusoids or wavelets, to encode a small number of bits of information in a pattern that

can be applied as a patch to an object. We first describe the SpEcBar for one-dimensional signals.

Consider a pattern defined by $f(x) = \sum c_i \varphi_i(x)$, where the $\varphi_i(x)$ are orthogonal basis functions, such as sinusoids with integral wave numbers, i.e., $\sin(nx)$, or wavelets centered at a particular location, $W_{2^i}(x - x_0)$, which is a wavelet of scale 2^i centered at x_0 .

If the pattern is analyzed by a decomposition with the basis functions, $\langle f, \varphi_i \rangle$, then the coefficients c_i are recovered. On the other hand, a naive observer viewing the pattern $f(x)$ will see a pattern with no particular significance.

Portions of the pattern may be obscured. With sinusoids, as long as the frequencies represented in the sum are sufficiently large, peaks in the spectral decomposition will be observed if enough cycles are represented in the observed portions of the pattern. Indeed, a sinusoid windowed by a function $w(x)$ results in a spectral response contribution of $\delta_\omega * \hat{w}$, which despite the blurring characteristics of \hat{w} , will still give a peak at ω . Normally, \hat{w} will be a sum of sinc functions, and thus will have a sizable peak at 0, and thus the shifted function $\hat{w}(\cdot - \omega)$ will have a strong peak at ω .

With wavelets, the same argument holds, but the spatial localization of the wavelet can be used to improve performance with respect to obscuration. Let $W_s(x)$ be a wavelet at scale s . The collection $\{W_{s_i}(x - x_{i,j})\}$ for a set of scales $\{s_i\}$ and for a set of spatial positions $\{x_{i,j}\}$ that depend on the scale s_i are orthonormal, and can thus form independent coding functions. Each function has essential spatial support that is localized around $x_{i,j}$. Thus the pattern $f(x) = \sum c_{i,j} W_{s_i}(x - x_{i,j})$ encodes the information in the coefficients $\{c_{i,j}\}$, and by using equal values for different spatial positions at the same scale, $c_{i,j} = c_{i,j'}$ for $j \neq j'$, a collection of coefficients are redundantly encoded by the pattern, and can be extracted by a wavelet decomposition, providing a sufficient portion of the pattern is unobscured. The set of frequencies encoded by the wavelets $\{s_i\}$ form a dyadic sequence, so that typically $s_i = 2^i$.

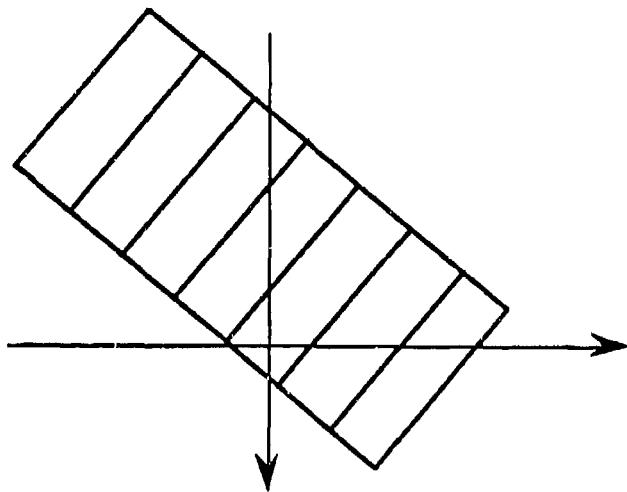
For a two-dimensional pattern, we use a one-dimensional SpEcBar and build a 2-D pattern by encoding the same information on each horizontal slice. One way to do this is to repeat the same function on each horizontal slice, creating a vertically striped SpEcBar. In this case, the 2-D pattern g will be given by $g(x,y) = f(x)$ in a rectangular patch.

However, there is no requirement that the SpEcBar be vertically constant. More generally, we may set

$$g(x, y) = \sum c_i(y) \cdot \phi_i(x),$$

where the $c_i(y)$ are constant for the particular indices i that encode the information that we wish to convey in the SpEcBar, i.e., $c_i(y) \equiv a_i$ for a subcollection of indices i . The coefficients that are not in the subcollection are allowed to vary. In either case, we will use $f(x)$ to denote an arbitrary horizontal slice through the SpEcBar, with the understanding that $f(x)$ may vary for different slices, but will always encode the same information.

In order to extract information from a two-dimensional SpEcBar, we require that a linear scan be obtained and digitized, where the scan intersects the SpEcBar at an angle no greater than 45° . This can be accomplished by scanning the scene in a raster-scan fashion twice, once from left-to-right and top-to-bottom, and the next time top-to-bottom and left-to-right. For each linear scan, we obtain a function $h(x)$ representing the scanline data for that slice.



We examine each linear scan to see if the SpEcBar code can be extracted from the corresponding function $h(x)$.

Since the scanline slice is not necessarily a horizontal slice through the SpEcBar, the function $h(x)$ will be a distorted version of the pattern $f(x)$. However, the distortion is necessarily a dilation:

$$h(x) = f(t \cdot x),$$

where t is a dilation factor related to the slope of the scanline through the SpEcBar.

Because of the dilation, the decomposition of h will not directly yield the desired coefficients. However, both the sinusoidal basis functions and the wavelet basis functions have the property that dilations of the basis functions are in the same class as the original set of basis functions, although they may not precisely equal one of the discrete collection of basis functions that are used in the orthonormal expansion. For sinusoids, we have that $\sin(t \cdot nx)$ is a sinusoid with wave number $n \cdot t$, and for wavelets, $W_s(t \cdot x) = W_{t,s}(x)$, which is a wavelet at scale $t \cdot s$. It can easily happen that the resulting wave number is non-integer or the scale factor for the wavelet is not a power of two. However, through numerical processing techniques it is nonetheless possible to extract the encoded information.

Although it is not necessarily the most efficient method, one way to handle the dilation is to simply decompose the observed scanline data $h(x)$ at various dilation scales, so that $h(x/t_k) = \sum b_{i,k} \varphi_i(x)$ yields a collection of coefficients $\{b_{j,k}\}$ for a set of candidate $\{t_k\}$. Because of the closed nature of the basis functions, it is not required that the exact value of t_k used in one of the candidate dilations equals the value of t representing the scan angle. Sinusoids are particularly advantageous here, since it is only required that t/t_k be close to integral.

The method of encoding information in a SpEcBar requires that:

- A SpEcBar code be recognized when it is seen;
- A SpEcBar code is not likely to be seen by chance;
- Some small amount of additional information is carried along with the SpEcBar code.

A method for accomplishing this using sinusoids would choose, say, four spectral bands at which a high spectral component is recorded, together with two other bands to encode information. The presence of large coefficients for the four bands encodes the presence of the SpEcBar. Neither phase nor magnitude information is important, save for large magnitudes of all four frequencies. A number of other designated spectral frequencies can then encode information. Both the amplitude and phase of the coefficients to these bands can, in a quantized fashion, yield information. The amplitudes can be measured relative to the amplitudes of the four identifying spectral components, and two bits of information can likely be carried in each amplitude. Likewise, the phase component can likely carry another two bits. Depending on the resolution of the system observing the

SpEcBar, encoded spectral lines might encode a sufficient number of bits to give a unique identifier from among millions of possible numbers.

With wavelets, the situation is slightly different—there is no phase information, but the coefficients can be more informative. However, because of the spatial redundancy, there is likely to be fewer available spectral lines to choose from, so that more bits should be carried on each channel. Information can be spatially encoded using the wavelet-based SpEcBar, although there is then potentially a loss of information due to partial obscuration.

Suppose that a total of 10 spectral lines carry information in a SpEcBar. In order to properly hide the SpEcBar in the non-designated lines, there should be many more available lines. For sinusoids, we might insist that there be 64 lines available, which means that 128 digitized samples should fall on the SpEcBar. If the SpEcBar is a small patch, then this means that highly accurate close-up observation is required. However, if the SpEcBar is large enough that the frequency components of the object of interest fall in the same range as the designated frequencies in the SpEcBar, then we can assert that the SpEcBar is readable whenever the object is discernible.

3. COMBAT IDENTIFICATION USING THE SPECBAR

For IFF (or Combat Identification) applications, the SpEcBar Code can be used to paint decals on a vehicle (or even a helmet) that can be observed, scanned, and analyzed in order to obtain an identification number. By using removable (or paste-on) decals, codes can be changed daily. The decals will have variable reflectivity patterns, which can be in the near IR wavelengths, or UV, or other spectral band that makes it less likely that the adversary will discover the code.

The interrogator scans and locates a decal, digitizes the reflectivity information, and processes the information using 1-D transformations. The scanning can be done numerically using a digitized array from a focal plane array, or can be done using a single-pixel sensor that scans the target using a raster scan or random linear scan fashion.

There are two possible modes of operation. In the Active SpEcBar system, the interrogator paints the target with a scanning beam and observes the reflection using a notch filter receiver. The spectral response of the reflected beam will be attenuated by the width of the spot size on the SpEcBar, so that the spot size must be kept small with respect to the designated encoding frequencies on the patch.

In the Passive SpEcBar system, the patch is observed using reflected background radiation, and a sensor/digitizer observes the object and performs numerical processing. In the passive system, the SpEcBar decal might well encode the information using variable IR reflectivity, so that the observing sensor is a FLIR (which can optionally be a one-bit gimbaled FLIR).

The required one-dimensional FFTs or wavelet decomposition's are easily performed in real time using DSP chips. Note that there is no requirement for a 2-D spectral decomposition. The only limiting requirement is that there must be a sufficient number of samples on the SpEcBar so as to both accurately locate the designated spectral bands, and so that the designated bands can be hidden in enough other lines so as to make the background marking inconspicuous.

**D. SOME COMMENTS ON AUTONOMOUS TARGET
RECOGNITION (ATR) RESEARCH**

Robert A. Hummel
Courant Institute of Mathematical Sciences
New York University
New York, New York

SOME COMMENTS ON AUTONOMOUS TARGET RECOGNITION (ATR) RESEARCH

INTRODUCTION

This study was conducted as part of the Defense Science Study Group, an ARPA-supported project to the Institute for Defense Analysis, by Robert Hummel. The study is informal and high-level, meant to provoke thought on the topic of Automatic Target Recognition as a defense science technology. We include a few technical suggestions concerning ATR, mostly as exemplars of the flexibility and range of research that can be conducted. The study is based on many briefings as part of the DSSG program, and a number of private briefings and discussions, acknowledged at the end of this document.

EXECUTIVE SUMMARY

ATR is possible, and should be developed, as it is highly worthwhile. ATR is not a problem that needs solving, but rather an array of technologies that can provide operational assistance in intelligence and combat situations. Development of ATR has been hampered by a lack of appropriate sensors, and a reluctance of developers and government sponsors to address intermediate goals and military relevance. Although there is good communication and cooperation between the services, technology transfer from pure research activities to system developers is poor; feedback as to the needs and constraints of the final system rarely happens; and the eventual performance of systems compared to the initial promise has been diminishing. Thus the research environment is not particularly healthy. Of the various approaches, model-based vision methods offers the best route, due to its reliance on simulation. Feature extraction and image representation based on features is the key to improved ATR performance. We offer some thoughts on image representation based on features, and using features to perform matching to models.

SOME BACKGROUND

Automatic Target Recognition (ATR) in a broad sense is the future of next generation weapons systems. The ability to perform complete ATR on objects in the combat environment using an autonomous system using non-cooperative sensing would

lead to complete situational awareness (SA), solve the combat identification (CID) problem, obviate identify friend-or-foe (IFF) systems, and imply large standoff ranges, as weapon systems could be made unmanned and brilliant. In a certain sense, ATR would provide an all-encompassing military situational analysis capability. ATR research in practice generally refers to image processing, but can encompass signal processing for object recognition. The objects are usually tanks, but each service has different needs. Nonetheless, the technology needs for object recognition and image and signal analysis cross many military domains and form the core of the ATR technology developments. ATR investment probably accounts for something like \$500 million per year in U.S. investment and is mostly basic research. Accordingly, much of the image processing research in the country is devoted to or motivated by ATR development.

ATR STUDIES

There was a 1983 Defense Science Board Task Force study of ATR technology and developments, whose conclusion was that the technology was not yet ready for an FSED phase. Among other recommendations, the study called for more data collection, and continued maturing of the field. A similar DSB report was issued in 1987, which saw some improvement in the situation, particularly with the emergence of model-based vision systems, but called for continued basic research. The LANTIRN program to procure a navigational pod and a targeting pod for the Air Force included specification of a target recognizer, which to date has not been selected for procurement, leading to an empty space in the targeting pod. Continuing programs investigate possibilities for the ATR component of LANTIRN. A recent JASON study in the summer of 1993 provides optimistic predictions, but calls for greater use of simulation, more consideration of countermeasures, and considers measures of performance and operating curve characteristics. Clearly, ATR is a large area in need of management and stewardship.

THE ATR CHALLENGE

One version of ATR concentrates on tank-detection from single-look, single-sensor images. Other variants include the use of multiple images, multiple sensors, and multiple target types, together with domain knowledge, all to enhance performance and achieve more realistic goals. Depending on the operational scenario, the challenge consists of detecting vehicles or regions of military significance, recognizing the threats, and/or classifying and identifying the targets. If carried out to the final stage, the goal has been to attach a target type to each detection, such as differentiating T72s from M1s, etc. In the post cold war, the situation is complicated by the fact that target type doesn't identify

ownership. A more practical definition of the ATR challenge views the needs for assisting in detection and military effectiveness as part of an overall situational awareness system, with a total system integration (potentially incorporating combat identification systems) required for operational evaluation.

For the purposes of research evaluation, current practice is to evaluate ATR performance using probability of detection PD and probability of false alarms PFA. Sometimes an entire operating curve (ROC) is plotted for a given algorithm. While it is commendable to use a common evaluation criterion, this particular method often bears little relation to the practical operational environment of an ATR and overemphasizes the autonomy of the system, neglecting rewards for partial but useful accomplishments. Specifications tend to compound the difficulty, especially by focusing on fixed target types, disregarding the need for robustness. A more realistic evaluation will require integration of an ATR algorithm into a more complete battlefield simulation system, which might include an IFF system, and include in the measurement of performance the adaptability to new targets and environments that might not even be suspected during the development process.

THE NEW ENVIRONMENT

As noted above, the new world environment changes the challenges for ATR, complicating the situation. Nevertheless, there has been renewed interest in ATR, with a proposed procurement expected for the Comanche helicopter, and many new basic research projects. What is driving this new interest?

One might argue that interest never waned, but development is getting new scrutiny. Certainly, Operation Desert Storm (ODS) highlighted the utility of high-technology, and the desirability of (1) image processing and image databasing, (2) autonomy in weapons systems for improved targeting accuracy, and (3) the need to identify quickly TELs in order to stop ballistic missile attacks before launch.

Another potential driving force for renewed interest in ATR is that it promises significant military advantages while anticipating low-cost procurements (relative, say, to production costs for bombers or missile defenses). Since development costs are likely to be a large portion of the procurement of computer systems (and especially computer software), investment of a few billion dollars might well lead to a system that can have a significant operational influence on a conflict that might cost (as ODS did, when both sides are included) tens or hundreds of billions of dollars. ATR also includes deterrence aspects. That is, ATR procurement provides the promise of considerable deterrence of future conflicts, by rendering ground armies more vulnerable.

Yet another motivation for ATR development is that we should not want the technology to be developed elsewhere without full understanding of the technology in the U.S. A surprise could be destabilizing, giving a potential opponent (possibly unfounded) confidence. Fortunately, ATR research is beginning to shed its defensive stance, with increasing confidence that a feasible system can be fielded in the foreseeable future.

ATR IS POSSIBLE

All previous studies have expressed some degree of optimism, but we state here more emphatically that ATR is possible and should yield useful, worthwhile systems within a couple of years. This statement is largely based on viewing ATR as a continuum of technologies so that partial "solutions" have operational merit. It is important to view ATR as (1) a force multiplier, where autonomous systems aid or cue personnel in order to reduce workload, and (2) a bandwidth compressor for communication, so that an ATR system is used to reduce the number of locations that must be subjected to analysis, permitting communication of imagery to more secure locales by unmanned systems. Only recently have systems attempted to make use of a second generation FLIR in conjunction with range or radar data, and although fusion methods are still largely research topics, they should make viable ATR far more practicable. Indeed, it is likely that much of ATR is easy.

Classical approaches to ATR fall into three categories: (1) matched filtering, (2) statistical pattern recognition, and (3) model-based vision. Later in this paper, we demonstrate equivalencies among the approaches but note that the form of representation of information is fundamentally different in each.

Matched filtering is equivalent to measuring a pixel-by-pixel difference between a stored model and the observed scene, in each local region over which the filter is applied. It is well known that this method does not work well when confronted with noise or real data. Using pixel data other than gray levels, such as edge information, can help some, but the ideal image representation information is not known. Matched filtering does not easily handle scale or rotation transformations.

Statistical pattern recognition views each location in the image as characterized by a vector $v = (v_1, \dots, v_n)$ that describes the "targetness" of the location, and then attempts to decompose \Re^n into classes representing different target and non-target types. By determining distributions, precise bounds on performance can be computed. The problems with this viewpoint include (1) large computational complexity ensues; (2) the vector must always be a fixed dimension, making it difficult to deal with feature components that are

defined only under certain conditions, as commonly occurs in image understanding where obscuration can occur; (3) the best dimensionality for the feature vector is unknown; (4) the design of feature components is critical and infinitely variable, etc.

Model-based vision uses models of the objects and the sensor systems to predict the observed characteristics of the targets and attempts to match observed patterns to predicted patterns. Model-based vision is universal, subsuming other approaches, since matched filtering and statistical pattern recognition also make use of models. However, model-based vision typically relies on geometric models for the targets and uses simulation and computing as opposed to empirical sampling for ATR development. Model-based vision turns the research from a statistical analysis of the variability of unspecified features into a directed research process of understanding and modeling targets, nontargets, and sensors and provides the only non-random research approach. It has led to considerable progress in academic research, as with the ACRONYM, SCERPO, Alignment, and Geometric Hashing systems.

PERSPECTIVES ON ATR

Although we view model-based vision as the appropriate route among the three classes of approaches, it must be developed in a very broad perspective. A particular problem for ATR development, as brought out by the JASON study, is that countermeasures can and will confound ATR algorithms, which is particularly problematic for model-based vision, in that explicit models of the intended targets are required. Likewise, camouflage poses difficulties, both in the detectability of targets and the increased variability of the targets. Countermeasures will typically add to the signature of the targets, although not necessarily in known ways. Panels that reflect or absorb IR radiation that can change internal detail as well as expand the shape of the targets mean that the signatures are more variable than typical ATR development assumes. Camouflage can modify the expected pattern of features, again in ways that might not be known in advance. Model-based vision approaches must include these variabilities in the invariance class over which recognition must be performed and might well require rapid adaptation of the target model database.

Inasmuch as ATR development is intended to provide technology that is useful for specific military needs, it is important that the technology that is developed is useful for a range of applications. Accordingly, it is not only the performance of an algorithm on specific tank and vehicle classes that matters, but also the programmability and adaptability of that algorithm to (1) new classes of targets and non-targets and (2) different classes of

transformations that cover the variations of single target types, while maintaining performance measures. In this way, methods that are developed can be useful for different military branches, regardless of whether the targets are tanks, aircraft, ships, submarines, or soldiers. Tri-service cooperation on ATR is high, but communication of those tri-service needs to contractors and researchers has not been sufficiently translated into modified evaluation criteria.

The JASON report suggests that hobbyists might begin developing ATR algorithms if data is made available. Additionally, there should be a greater civilian technology pull for ATR-type technologies.

Medical image processing, a developing field for analysis and diagnosis of medical sensed data for clinical applications, provides one possible domain for a greater civilian involvement. While the application is different, the techniques and goals overlap in critical ways, such as in the analysis of multidimensional multimodal data to discover anomalies or target characteristics. ARPA should encourage the development of medical imaging applications for pre-surgical planning, diagnosis, and surgical assistance, as a means of providing impetus to the commercial development of technologies that are useful for ATR, as well as providing for techniques that could have considerable military significance in providing medical care.

A difficulty in ATR development, noted as early as the 1983 DSB report, is that the community has been reluctant to develop partial solutions. For example, improving sensor resolutions implies that algorithms that require relatively many "pixels on target" today will provide greater utility tomorrow. Often, however, specifications have thwarted research by requiring that algorithms address the problem of recognizing tanks at 5 km with fewer than 10 by 10 pixels expected to fall on any single target. Since ATR is a range of technologies, involving cueing, filtering, prioritizing, and refining (as in a situational awareness map), algorithms must be permitted to be developed according to stages of utility, with practical horizons from each stage. For example, an algorithm that can reliably detect tanks from a narrow field-of-view close-range sensor (thus providing many pixels on the target), when combined with an algorithm that can detect moving objects in a dynamic scene that is possibly subjected to optical flow due to motion of the sensor, together with image stabilization capabilities (as developed already for camcorders), can lead to a detect/look/stare system using a gimbaled multiresolution sensor. (Indeed, programs exist with each such component, with this kind of system in mind, but also incorporating multisensor information.) In such a system, even the narrow field-of-view recognition

system, in isolation, can provide operational utility in terms of refining aim points or providing an autonomous land vehicle with added capabilities.

The main perspective on ATR that we state here is that matched filtering, statistical pattern recognition, and even model-based vision are all excessively rigid for ATR technology development, and that new algorithmic means are required. The new methods are likely to be most closely allied to model-based vision, but will not likely use CAD models of the three-dimensional shapes and structures of the target classes, and match observed patterns to perspective-invariant projections. The ATR "problem" is broader and more general than that.

Indeed, the key to the new algorithmic methods will be the features that are extracted from the data. Matching of patterns of features is important, but is relatively easy compared to the question of which features best characterize the image to assist in recognition of targets and to separate targets from the other objects in the scene. In the past, features have often been extracted from target segments. The target segments attempted to classify pixels as to falling on the target region or off of the target region. While this may have made sense for first generation FLIR sensors that provided little internal detail of the targets and relatively few pixels on the target, we will argue here that features should be "free-floating" and indicative of image characteristics and not tied to a presegmented region of interest.

RESEARCH ENVIRONMENT

We identify four classes of players in the ATR research community: (1) university researchers, largely associated with the ARPA Image Understanding program (although increasingly supplemented by other programs at ARPA); (2) research houses, including both FFRDCs and private shops, often employing university graduates; (3) government labs, such as NVL, ARL, NAWC, MICOM, AFAL, and others, providing the glue between the players, and (4) industrial research shops, where the major players include Martin Marietta, Hughes, Westinghouse, Honeywell, and others.

Each class can be criticized, but overall, each class plays a vital role, and within each class, cooperation and teaming are good. (This is a recent change, especially for industry, where proprietariness used to be rampant.) The ATR Working Group (ATRWG) "society" has helped in this regard. However, there is less success in fostering technology transfer between the classes. Each class seems inclined to want to repeat the roles of the other classes, in some cases because they do not believe that the other class is fulfilling its role. As an example, university researchers do not seem to understand or care much about

the military context of the ATR development, nor any other particular application domain.¹ Likewise, the industries tend to want to develop their own algorithmic methods or to blindly implement a published algorithm without suitable adaptation and consideration for the particular sensors and scenario, thereby locking in ancient technology. To a certain extent, these difficulties are unavoidable, and ATR development is likely to progress regardless, but the research environment in ATR seems particularly unhealthy.

ALGORITHM DEVELOPMENT IN DOD

With the emergence of information technologies as a critical defense science domain, it is important that defense procurement understand algorithm and software development. Evidence, especially in the ATR development area, is that this level of understanding is low. Too often, software procurement is regarded in a similar fashion to radar system development: hardware modules should be developed and plugged in according to schedule. In ATR, the term "algorithm" refers to a technique or approach and not necessarily to a recipe. When an algorithm is described that does refer to a recipe, such as an edge operator or a Hough transform, the algorithm is often used to solve an imprecise problem. The algorithms are like proofs without the theorems. Nearly all computer scientists have had the experience of spending weeks on some code, only to realize that the data structure is wrong or that the code doesn't implement the algorithm properly or that the algorithm is wrong, and that the entire code must be discarded. Programmers sometimes have the same experience, but for different reasons. An algorithmicist will understand that the difficulty in algorithm development is to formulate the problem in mathematical terms, such that the solution can be found by efficient processing methods. It is the formulation of the problem and the representation of the information that is the principal challenge. For ATR development, this is especially difficult, since the goals of ATR can be varied and wide-ranging.

The algorithms, meaning approaches, that are used for ATR development, suffer from a "mathematical fad of the week" syndrome. It is irrelevant who is to blame for this phenomena. The protagonists feed off one another. The academic IU community contributes in that they develop buzzword ideas and communicate with one another and with industries using terms that hide the operational idea. In some cases, this has a beneficial side. However, it also contributes to the perception that the algorithm issue is a matter of choosing some mathematical system, as opposed to defining the problem. A

¹ The DSSG program supports a redress of this criticism.

prototypical example is the neural network approach, which is touted as an algorithmic method that obviates thinking. In fact, there are many interesting and useful methods that can be described as neural net approaches. However, saying that one can apply neural net methods to solve the pattern recognition problem to produce improved ATR sensors misses the essence of the problem as a feature extraction and pattern matching system.

UNIVERSALITY OF MATCHED FILTERING

Matched filtering can be formulated as the task of maximizing a collection of vector dot products, so that if \mathbf{x} is a vector representing the ordered pixel values, and $\{\mathbf{x}_i\}_{i=1}^n$ is a collection of target vectors, then the problem is to find the index that maximizes $\mathbf{x} \cdot \mathbf{x}_i$ over all possible i . Since the \mathbf{x}_i will typically contain multiple translates of the same prototype pattern, many of the vector dot products can be efficiently implemented as convolutions. Since the \mathbf{x}_i are not orthogonal, there can be considerable cross-talk. The vector \mathbf{x} of pixel values may represent the result of an edge detector or other filter of raw sensor data, and may also incorporate multispectral and multisensor information.

Although there are many forms of statistical pattern recognition, a K-nearest neighbor classifier can be used to arbitrarily closely approximate any desired distribution in feature space, which in turn can be arbitrarily closely approximated by a one-nearest neighbor classifier (albeit with many more prototypes). In this classifier, a feature vector \mathbf{x} locates the prototype that is nearest, i.e., minimizes $|\mathbf{x} - \mathbf{x}_i|^2$ over i . This is equivalent to maximizing $\mathbf{x} \cdot \mathbf{x}_i - \frac{1}{2} |\mathbf{x}_i|^2$, which we may write as $(\mathbf{x}, 1) \cdot (\mathbf{x}_i, b_i)$, where $b_i = -\frac{1}{2} |\mathbf{x}_i|^2$. Thus we once again get a vector dot product maximization, but in this case, the vector represents the feature values of a region of interest with an appended component that is always one, and the matched filters encode the prototype feature vectors, with an appended bias terms.

Model-based vision methods may be viewed as variations on the following. Features are extracted from the scene, and a basis set of features is chosen, which is a minimal set of features sufficient to determine a transformation to a corresponding basis set in a model. We must either find the corresponding model and basis set in that model by means of a search, or reject the hypothesis, based on the other features in the scene. If the hypothesis is rejected, then a different basis is chosen, and the process iterates. The features in the scene are normalized with respect to the basis set, forming a collection of points in a Euclidean space, which we model as a sum of delta functions $f(x) = \sum_i \delta(x - y_i)$.

The domain will be two-dimensional if the features are points in \Re^2 , but can have more dimensions if the features have attributes (as described in the next section). Generalizations using distributions that model the uncertainty in the measurement of the features is possible. The scene $f(x)$ is compared to large collection of normalized patterns of models. For every model m and every basis B that can be chosen in the model, there is a normalized pattern, $g_{m,B}(x)$ that encodes not only the positions $\{y_{m,B,i}\}_{i=1}^{n_{m,B}}$ of the normalized features as in f , but will be generalized to include a distribution around each normalized feature representing the expected covariance $C_{m,B,i}$ about each feature. The distribution is used so that each corresponding feature in f can contribute a weighted vote toward the model/basis pair m, B according to how closely the feature in f corresponds to the feature in $g_{m,B}(x)$. The precise formula used to encode the distributions depends on the computational method that is developed for the recognition algorithm. Accordingly, finding the best model/basis pair can be viewed as maximizing $\int f(x)g_{m,B}(x)dx$ over all possible m, B . Since an integral vector product can be viewed as a vector dot product if the domain is finely sampled, we once again see that recognition becomes tantamount to maximizing a dot product.

Of course, in each of the methods, algorithmic variations can enhance the efficiency of the search.

Although each of the classical methods for object recognition can be formulated as a vector product maximization problem, the methods have an important difference. The representation of the information is different in each. For matched filtering, the information x encodes pixel data; for statistical pattern recognition, x is a feature vector describing a segment; for model-based vision, the description $f(x)$ encodes a transformation-invariant version of the pattern of features extracted from the scene for matching to expected patterns $g_{m,B}(x)$, where the features are normally edge elements, corners, and other geometric attributes.

The conclusion is not that matched filtering is the best approach, but rather that the representation of the information is the principal design question.

ATTRIBUTED FEATURES FOR ATR

What features are best to represent the information in the sensed data so as to facilitate recognition of targets? This is the main research question, but we offer two observations (without much explanatory detail).

First, individual features should be descriptive and "free-floating," rather than being based on "segments" consisting of binary regions extracted from the scenes. Feature extraction is a means of segmentation, but too often features are extracted through a two-step process, where first a region consisting of a connected set of pixels is designated as a region of interest, and then some measure is applied to the region, possibly using the pixel values within the region. Although a staged process of this type can help stabilize the feature extraction process, it makes it difficult to analyze the precise meaning of a given feature. When sensors provided little internal detail to potential targets, and the silhouette of the target was the main identifying characteristic, feature extraction based on binary segments made better sense. With the advent of better sensors and more internal detail, edge pixels are not necessarily expressive enough (by themselves, even in combination). Instead, features such as corners, intersections, parallel edges, lines, end-stoppings, etc., might well become important. Interestingly, many of these features are also known to be important in early stages of human visual processing, and guidance from neurophysiology and neuroscience experiments could be very important in determining image features. Note also that these features are attributed, in the sense that they have position information as well as other identifying information, such as orientation (as of an edge or line, or even of a corner), scale, and opening angle (such as with a corner or intersection). Accordingly, with rich attributed features, the vector representing the feature will typically have components representing not only the location of the feature in the scene, but also the attribute values.

As a simple example of an attributed "free-floating" feature, consider a simple corner detector. If the image data is given by $I(x,y)$, and an edge is defined to locally an isocontour where $\nabla I(x,y)$ is both large and maximal (in the direction of the gradient), then the local curvature can be seen to equal to:

$$\kappa = \frac{I_{xx} I_y^2 - 2 I_x I_y I_{xy} + I_{yy} I_x^2}{|\nabla I|^{3/2}}$$

If we compute κ at those locations where there is an edge, and perform non-maxima suppression, then we obtain points of high curvature along edges, which is equivalent to a corner. Note that this kind of feature extraction does not depend upon first identifying and extracting edge contours, other than using an isocontour model for edges.

The second observation that we wish to make is that features should be independent. By this, we mean that a feature vector v as extracted and described above should give no specific evidence as to the likelihood of the components of any other feature

vector. For example, corner information, where the position of the corner and orientation of the bisector of the corner is given, provides independent information. Edge elements with orientation, however, do not, since a horizontal edge very likely indicates the presence of horizontal edges in positions that are horizontally displaced from the position of the element. The reason independent features are important is that methods for accumulating evidence based on the features typically assumes independence. (Certain iterative methods, such as Markov Random Field models, provide exceptions.) Generally, if a collection of features contain dependencies, then they should be grouped and re-represented by a construct that provides independent information about the coalesced structure.

There are many possibilities for different features that conform to these observations; the use of sophisticated feature types for object recognition and ATR algorithms is in its infancy.

SOME SPECIFIC ALGORITHMIC IDEAS

As a way of indicating some of the algorithmic possibilites that need further exploration, we discuss briefly two ideas that can be applied, respectively, to feature extraction and pattern matching.

Feature-based image decomposition

Historically, feature extraction begins with an image I , applies various inner products or convolutions, as in $c_i = \langle I, \varphi_i \rangle$, and then performs some nonlinear operations in order to extract a collection of coefficients. If the φ_i are orthonormal, then one has a Fourier decomposition (related to a Fourier transform, but more general), with $I = \sum c_i \varphi_i + R$, (R is a remainder), and $|I|^2 = \sum c_i^2 + |R|^2$ (the Parseval Theorem).

Suppose we instead have multiple collections of filters, $\{\varphi_{k,i}\}_{i=1}^{n_k}$, for $k = 1, \dots, m$.

We may then develop the following *feature-based image decomposition* method. We set $I = I_0$ and $k = 0$. We apply the k -th bank of filters:

$$c_{k,i} = \langle I_k, \varphi_{k,i} \rangle$$

We then apply non-maxima suppression and any other nonlinear processing so as to suppress many of the $c_{k,i}$ coefficients (over i), to obtain a collection of coefficients $\{a_{k,i}\}_{i=1}^{n_k}$ such that each $a_{k,i}$ is either zero or is equal to $c_{k,i}$. We then set

$$I_{k+1} = I_k - \sum_{i=1}^{n_k} a_{k,i} \varphi_{k,i}$$

This process iterates for $k = 0, \dots, m$.

A remarkable thing occurs. Even if the $\{\varphi_{k,i}\}_{i=1}^{n_k}$ are not orthonormal, we still get a Parseval relation, providing the nonmaxima suppression ensures that only one nonzero coefficient survives in the region of support of any given $\varphi_{k,i}$. That is, providing $a_{k,i} \neq 0$, $a_{k,j} \neq 0, i \neq j$ implies that the support of $\varphi_{k,i}$ and $\varphi_{k,j}$ are distinct, then

$$\|I\|^2 = \sum_{k,i} a_{k,i}^2 + \|I_{m+1}\|^2$$

The relation occurs because each successive projection of I is an orthogonal projection, even though the collection of basis functions are not necessarily mutually orthogonal. The net result is that the image I is decomposed into features, according to $I = \sum_{k,i} a_{k,i} \varphi_{k,i} + I_{m+1}$, just as in a Fourier decomposition.

Feature-based image decomposition allows one to consider feature extraction where, for example, corners are detected and removed, then edges are removed from the result, and finally blobs or points are detected and removed from the residual. The advantage of such an approach is that edge detection, for example, can be considerably easier when corners are removed first.

Of course, the technique tells us nothing about the appropriate ordering of feature extraction, or even what kinds of features should be extracted. Feature-based image decomposition only increases the range of possibilities.

Geometric hashing for pattern matching

Geometric hashing is a method for organizing pattern matching and should be applicable to ATR research. Geometric hashing as a search method is particularly attractive because (1) it is parallelizable, (2) efficient especially when dealing with large databases, and (3) permits easy adaptation.

Using independent features with attributes, an image scene is represented as a collection of vectors $\{v_i\}_{i=1}^s$. For ATR applications, recognition must be translation invariant, and potentially invariant to some degree of scale change (when range is uncertain), and some degree of rotation (due to inclination variations of the target). In this

nominal description, viewer-centered models are required of all targets (meaning a model for every significantly different possible viewing direction). Each such model m is represented by a pattern of feature vectors $\{ \mathbf{u}_{m,i} \}_{i=1}^{n_m}$.

We define a basis set to be a subcollection of vectors such that a correspondence of one basis set to another specifies a transformation under which recognition should be possible. We pre-store all models with respect to every possible choice of basis set. This is done as follows. A hash function $h(\mathbf{x}_1, \dots, \mathbf{x}_{q+1})$ is chosen so that the result lies in a Euclidean space \mathbb{R}^p and is invariant (or nearly invariant) to uniform transformation of the features. For each model m , for every basis $B = (\mathbf{v}_1, \dots, \mathbf{v}_q)$ in m , a collection of "entries" in \mathbb{R}^p is obtained at locations $\omega_{m,B,i} = h(B, \mathbf{u}_{m,i})$ for every $\mathbf{u}_{m,i}$ in the model m , with "tags" that point to the model/basis pair (m, B) . Note that for every model, for every basis, there is a collection of entries that provides a normalized pattern representing the entire model (see the accompanying figure).

The geometric hashing algorithm begins by choosing a basis set B' in the scene $\{ \mathbf{v}_i \}_{i=1}^s$. The hash function is then applied to tuples to obtain points $h(B', \mathbf{v}_i)$ in the hash space, for each i . For each such hash point, entries are located that lie nearby (using, for example, a hashing scheme to bins that are used to discretize the hash space, or using a k -D algorithm, i.e., a multidimensional binary search tree), and each such entry $\omega_{m,B,i}$ registers a "vote" for model/basis pair (m, B) . The vote can be weighted: a large vote is registered if $h(B', \mathbf{v}_i)$ lies very near $\omega_{m,B,i}$, and the vote is weaker as the distance increases. In this way, the pattern created by the collection $\{ h(B', \mathbf{v}_i) \}_{i=1}^s$ is compared to the pattern created by $\{ \omega_{m,B,i} \}_{i=1}^{n_m}$ simultaneously for every m and B . Model/basis pairs receiving large cumulative votes should then be tested more carefully for matches with scene features. The precise formula that is used to measure "nearness" and the details of the weighted voting, accumulation, and verification schemes are issues for further research.

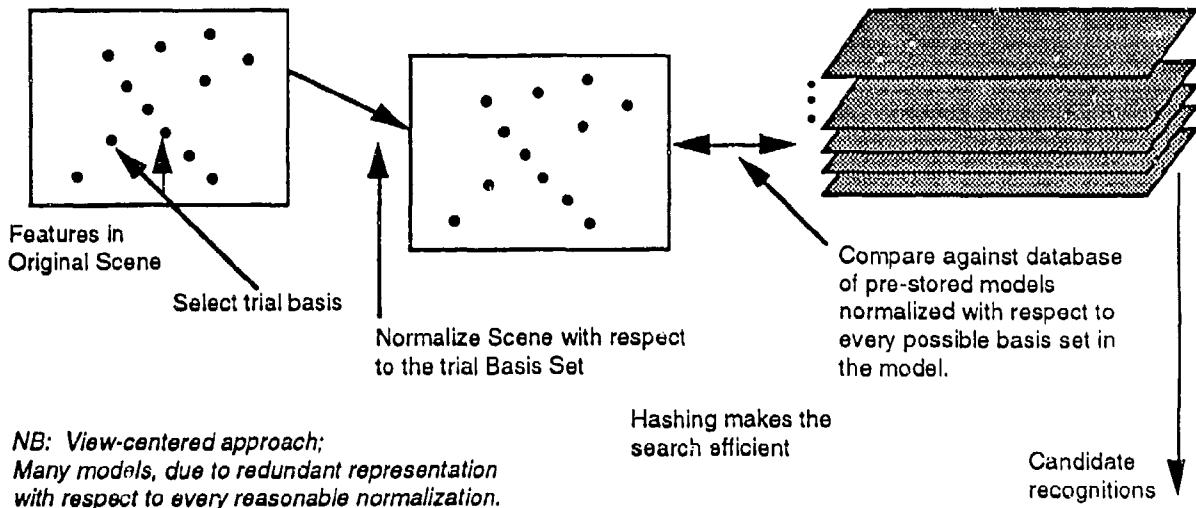


Figure 1. A Depiction of the Principal Steps in the Geometric Hashing Approach to Object Recognition

While geometric hashing provides a straightforward way of organizing pattern matching (and, as we have observed, is equivalent in some form to matched-filtering), it leaves open many issues, especially concerning the features and hash functions used to represent the objects. Geometric hashing has been studied in academic environments for a number of years, but has not been applied, in any concerted fashion, to ATR research. We conclude that ATR research has yet to tap an array of algorithmic methods that can provide important structural discipline to the development of sophisticated and successful systems.

ACKNOWLEDGMENTS

The DSSG was briefed on ATR at numerous places, including NVL, Martin Marietta Denver, Lincoln Labs, MITRE, and other places. Dr. Oscar Firschein and James Crowley at APPA fund ATR work and have recently granted funding to NYU to pursue an NYU-developed object recognition method (called Geometric Hashing) in conjunction with wavelet methods for feature extraction for ATR work. At IDA, Jeff Nicoll and David Sparrow have been generous in providing information and perspectives on ATR work. At Los Alamos, Greg Canavan and Chris Barrett provided extremely useful information. Dr. Peter Weinberger provided details about the JASON report, and conversations with David Reade were very helpful. I especially wish to thank Mark Hamilton, Lynn Garn, and Ed Zelnio, for a detailed high-level private briefing in Washington, D.C., in October 1993.

E. LATERAL WAVE MODIFICATIONS FOR ELECTROMAGNETIC PROPAGATION NEAR INTERFACES

**The Implications of a Theoretical Advance in Electromagnetic Signal
Propagation for the Design and Operation of Radar Systems**

**S. James Gates, Jr.
University of Maryland
College Park, Maryland**

LATERAL WAVE MODIFICATIONS FOR ELECTROMAGNETIC PROPAGATION NEAR INTERFACES

1. INTRODUCTION

The physics of an electromagnetic wave propagating at or near the boundary of two regions that have very different dielectric and/or susceptibility properties is important in maximizing the performance of radio and radar systems. Indeed, this is the usual situation that one finds on considering the possibilities associated with radio waves traveling near the surface of the earth. It might be expected that some unusual and unexpected phenomena arise.

One of these is the notion that at the interface of two rather different media, electromagnetic radiation can be guided in a manner similar to that which occurs in a wave guide. Stated most simply, there is almost an expectation that a radio signal broadcast near the surface between air and water will propagate along the air-water interface in a manner that is very different from a signal traveling either upward into the air or downward into the water.

This document reports on some of the implications of the existence of this closed-form, analytical lateral wave solution for the theory and application of radar detection and radio communication. Some possible implications for GPR (ground penetrating radar) OTH (Over-the-Horizon) radar systems are noted.

2. LATERAL WAVE CONCEPT

It might be expected that such behavior has by now analytically, phenomenologically and numerically been well understood. The topic of the propagation of electromagnetic dipole radiation near such interfaces is a very old one, after all. This effect was first explained by Zenneck in 1907 [1] who noted this type of behavior can occur in solutions of Maxwell's Equations that describe a surface wave guided along the interface of two media with different electromagnetic properties. At that time, the ionosphere was not known to exist and Zenneck's work was an attempt to explain the phenomenon of short wave radio transmission. Zenneck's solution required the existence of a so-called "lateral wave" whose characteristic feature is a decay rate of the electromagnetic field that is substantially

slower than one might naively expect. The electric field \mathcal{E} , for example, was predicted in certain circumstances to have a form,

$$\mathcal{E} = \mathcal{E}_0 \left\{ \frac{1}{\sqrt{D}} + \dots \right\} .$$

where D is the projection of the distance (in dimensionless units) from the dipole along the interface. The startling feature of this result is that naive expectation would lead to,

$$\mathcal{E} = \mathcal{E}_0 \left\{ \frac{1}{D} + \dots \right\} .$$

Since the intensity of the signal goes as the square of the field intensity, it is clear that a lateral wave is detectable along the interface, in principle, at distances much farther than the usual dipole wave that forms the principle component of most antenna radiation patterns.

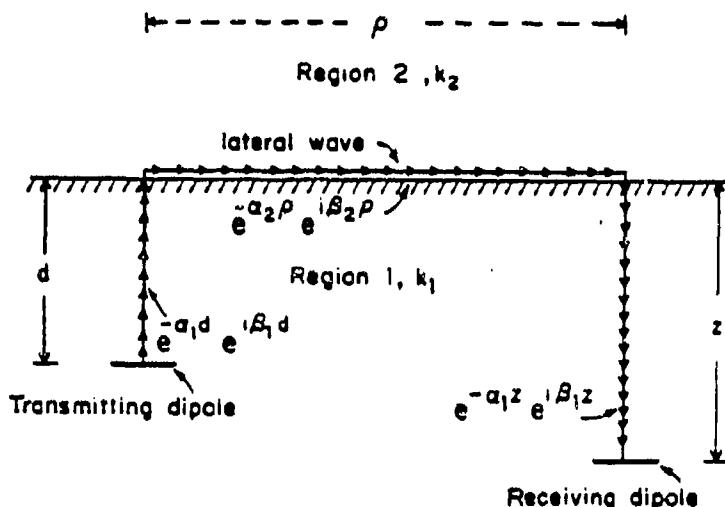


Figure 1. Lateral Wave Transmission Between Buried Dipoles

A classical work on the subject in 1909 by Sommerfeld [2] derived integral expressions for the electromagnetic field generated by a vertical dipole near the boundary of two different media. The dipole is the simplest possible model of a radio antenna. He also showed that a wave of the type envisioned by Zenneck emanates from an oscillating dipole placed at the interface between a dielectric and an imperfectly conducting plane. However, shortly thereafter, he repudiated this result and there came about a long period in which the existence or nonexistence of lateral waves associated with radiating dipoles near boundaries was a subject of disagreement.

Following this extended period of controversy about the correctness of such mathematical solutions, in more recent times the complete analytical solution for the case of the oscillating (radiating) point dipole near an interface has been derived by Wu and King [3,4] in a form that is both explicit and useful. These authors brought to bare a new technique of applying Fourier transformations to the boundary value problem. They were able to derive results that were free of the approximation most frequently used to evaluate the integral formulae of Sommerfeld. The naive expectation turned out to be true, a wave of electromagnetic energy tends to travel along the surface.

The Wu and King solution shows precisely the kind of behavior noted by Zenneck. The distinguishing feature of these lateral waves (or Norton surface waves as they are called depending on the location of the source relative to the interface) is that their amplitudes as a function of distance from the source is much larger along the surface of the interface than perpendicular to the interface. They are more robust and offer unexpected possibilities for detection, only along the interface, than might be expected for a pure dipole in free space. The proper incorporation of the result of Wu and King into the theoretical analysis of radar and radio communication operation was begun sometime ago [5] and continues even presently [6]. Recently, a study of implications of the lateral wave contribution to the low-altitude radar propagation factor was given [7]. As pointed out in this last reference, the expense as well as the effort directed toward target recognition algorithm development in the case of these systems suggest that the analytical solution of Wu and King should be used wherever it may be applicable. This is particularly important since much of the analysis of such systems continues to follow the usual albeit traditional analysis.

3. LW & RADAR PROPAGATION

The basic feature of radar returns in free space are governed by what is called the "radar equation." This simple equation basically describes the power of the returning signal that a radar transmits. Following Nathanson [8]¹, this takes the form:

$$P_r = P_t \frac{G_a G_t \sigma \lambda^2}{(4\pi)^3 R^4} .$$

¹ Skolnick [9] defines a slightly modified radar equation adapted to OTH radar. For our purposes, the simple form above suffices.

This equation is a standard tool of the radar design engineer and provides a useful theoretical limit on the operating efficiency of a radar. (In writing this equation, we have chosen its simplest form, thus neglecting a number of "loss factors." A standard reference where these are discussed is the *Radar Handbook* by Skolnick [9].) As can be seen, one of the stringent limits implied by this equation is that the loss in the returning signal goes like the inverse fourth power of the distance! However, this "R⁻⁴ loss" has a simple and, for our purposes, interesting interpretation. We can write:

$$\frac{1}{R^4} = \frac{1}{R^2} \frac{1}{R^2}$$

and the interesting feature is that each of these factors of $\frac{1}{R^2}$ can be associated with the intensity of the outgoing and returning beams, respectively, of electromagnetic radiation. The fact that these go as the inverse square of the distance is characteristic of dipole radiation in free space. This is where the observation of lateral waves is so important. Since the intensity of lateral wave propagation goes as $\frac{1}{R}$ not $\frac{1}{R^2}$, the most naive calculation would imply that the radar equation must be modified to replace the "R⁻⁴ loss" by a "R⁻² loss!"

It has been suggested [7] that the role of lateral waves can be accounted as "multi-path effect." In multi-path, the path of the signal (either outgoing or returning) can be viewed as possibly having one "bounce off" the surface of the earth. The multi-path correction to the radar equation appears as a modification to of the basic radar equation

$$P_r = P_0 \frac{G_a G_t \sigma \lambda^2}{(4\pi)^3 R^4} |F_{a \rightarrow t} F_{t \rightarrow a}|^2$$

where $F_{a \rightarrow t} (F_{t \rightarrow a})$ is known as the "antenna-to-target (target-to-antenna) pattern propagation factor." The interpretation of these are

$$F_{x \rightarrow y} \equiv \frac{\text{"the true electric field at } y \text{ due a source at } x\text{"}}{\text{"the naive electric field at } y \text{ due a point source at } x\text{"}}$$

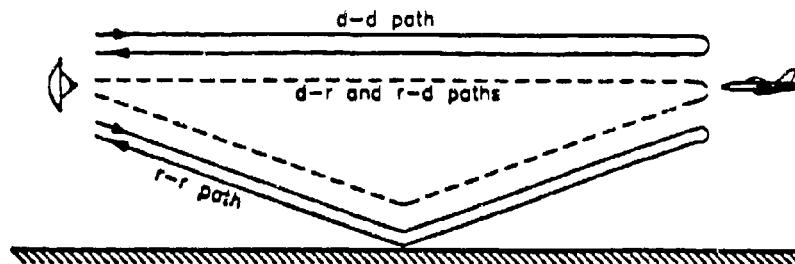


Fig. 2. Diagrammatic Indication of Different Paths Contributing to Multi-path Corrections

In traditional analyses this is calculated in the so-called Fresnel approximation. In this analysis the polarization-orientation (i.e., vertical or horizontal) of the radiation is important, but we will neglect that for now. The traditional analysis involves using the Fresnel approximation and geometric optics. Since the Wu-King solution is an exact analytical solution, it can be compared to the Fresnel approximation and the ratio given by

$$\eta_{LW} = \left| \frac{F^{\text{total}}}{F^{\text{Fresnel}}} \right|$$

is the correction to the radar equation due to lateral waves.

$$P_r = P_0 \frac{G_a G_t \sigma \lambda^2}{(4\pi)^3 R^4} |F_{a \rightarrow t} F_{t \rightarrow a}|^2 (\eta_{LW})^4 .$$

As a measure of what this can do to the performance of a system, we can use the maximum range ρ_{\max} for a given signal-to-noise ratio (S/N) as well as bandwidth $\Delta\omega$ and temperature T to find

$$\rho_{LW-\max} = \sqrt[4]{P_0 \frac{G_a G_t \sigma \lambda^2}{(4\pi)^3 (S/N) k T (\Delta\omega / 2\pi)}} \sqrt{|F_{a \rightarrow t} F_{t \rightarrow a}|} (\eta_{LW})$$

For an appropriate range of conditions we see that, theoretically, lateral waves predict an unexpected increase in performance.

4. LW & INTERFERENCE PATTERNS

There is another way in which LW modification can have striking implications for the use of radar. With the onset of synthetic aperture radar (SAR) [9,10], the possibility of using the phase data contained in the interference pattern of a returning signal suggests that there is a second possible application that needs to be thoroughly studied. In two of the included appendixes, some sample expressions for LW and ordinary dipole radiation components of the electromagnetic field are given. As can be noted these expressions are rather different. The presence of the lateral wave terms (proportional to \mathcal{F}) allows for new interference patterns of the electromagnetic field that are not present without the LW contributions. There has been some reported success in experimental work done to probe for the existence of these LW-signature interference patterns [11]. Recognition of these patterns opens the way to write specialized signal-processing codes that could enhance radar detection technology. The mathematical analysis of this mode of exploiting LW radiation is dependent very much on a detailed mathematical analysis of the phase history of an electromagnetic beam as it is detected along a moving path. We will not undertake here to give an explicit example.

5. APPLICATIONS

Before looking at issues raised by the existence of the closed-form, analytical Wu-King solution, it is probably appropriate to ask what comparative advantage there might be to making a call to support a large-scale reexamination of issues that many regard as being settled. The main criticism one might raise is, "Although the Wu-King solution may be correct, the advances in the field of understanding and designing radar systems has been so spectacular, their result cannot have very broad impact?" This criticism may be valid, but the existence of this solution points out that there are features of electromagnetic propagation that are subtle enough to have been missed for long periods of time. Because the effects must be subtle, they may open opportunities that might have easily been missed had one not known of their *a priori* possible occurrence. The challenge is to exploit this subtlety into a highly leveraged benefit. Since the studies at this point are all purely formal, the cost of such an effort is very moderate. Taking a chance on such a small investment surely is prudent.

Along this direction, it is useful to see how the Wu-King methodology manages to be effective within the confines of such mature and well developed field. This can act as an intermediate step to bolstering arguments for a broader study. One way to do this is to compare to it to competing approaches.

A. Comparison of Standard and Wu-King Methods

The key to understanding what is the significant new advantage of the method of Wu and King over the standard approach to the problem of the radiating dipole new boundaries can be seen by noting the difference as compared to the standard methodology. The rules for the standard methodology were established by Sommerfeld [2] who was the first to establish a set of integrals to represent the electromagnetic fields due to a dipole near an interface. We illustrate these following the presentation of Kong [12] which is equivalent to that of Sommerfeld.² The basic idea is to use the Hertz potential for the calculation of the electromagnetic field due to a dipole near a boundary. This leads to the field components being represented by a contour integral. For example the H_z component of the magnetic field takes the form (See Annex A, Section 3 for a more explicit presentation)

$$H_z = \int_{SIP} d\xi F(\xi)$$

This contour integral is to be taken over the "Sommerfeld Integral Path" which contributes to some difficulty to abstract results that are useful to the engineer. In fact this one is a major contribution of the work of Norton.

The key to understanding the method of Wu and King is to note that instead of representing the exact solution totally in terms of contour integrals, they observed that it is possible to Fourier transform the boundary value problem, but only parallel to the direction of the interface. This typically leads to some simple differential equations involving the direction perpendicular to the interface. These are solved explicitly and then inverse Fourier transforms are used to calculate the spatial dependence with respect to the coordinates parallel to the interface. This simple idea has given a surprisingly powerful tool for evaluating this class of problems. The advantages arise due the fact that fewer approximations are necessary to evaluate the inverse Fourier transform integrals.

B. Comparison of Wu-King Methods and NEC

Since there have been available for some time high-speed numerical calculations [14], one might suggest that an extensive investigation of the implication of lateral wave propagation is unnecessary. For example, there exists the National Electromagnetic Code (NEC) for calculating field strengths in various situations. But even here there are certain

² See also King and Smith [12].

cautions that should be observed. Along these lines, the following quote from King, Owens and Wu (1993) is notable.

Although numerical methods and high-speed computers have proved to be very valuable tools in the evaluation of fields that depend on the complex integrals of Sommerfeld, they do have serious limitations. These are a consequence of the unavailability of analytical expressions that give insight into and understanding of the complicated fields. Even very extensive and systematic evaluations and tabulations like those of King and Smith (1981) failed to reveal the physically fascinating and important interference patterns generated by the surface wave and direct wave near a horizontal dipole in various half-spaces bounded by air. These are described and discussed later. Similarly, the numerical calculations of Young and Cox (1981), while correctly calculating the desired specific electric field on the sea bottom, failed to disclose the reasons for the very different behavior of the components or the orientation of the antenna required for simple physical interpretation of the results. This is also discussed later. Computer-based numerical evaluations of the field of dipoles near an interface are powerful but essentially blind. They are a most valuable supplement to, but no substitute for, analytical formulas when these are both accurate and conceptually simple. It is the purpose of this book to derive and apply such formulas subject only to the general restrictions (1.4.1)—or in some cases only—which are almost always satisfied in practice.

C. Implications for OTH Radar Studies

Having seen how the existence of lateral waves can have dramatically change theoretical predictions of the radar equation, it is perhaps not too fanciful to conjecture what might arise from their inclusion into considerations of the operation of Over-The-Horizon (OTH) radar systems.

It can be seen from the nature of the lateral wave that there are two conceivable ways in which lateral waves might modify our mathematical understanding of OTH systems. These two ways are intrinsically tied to the two operational modes of constructing OTH systems. One of these is known as the "skywave" mode and the other is known as the "groundwave" mode [9,15]. Within this arena, an important role is played by the ionosphere. This suggest that a study of the implication of lateral waves for OTH radar can be posed within the context of a multi-layer three Wu-King solution. Such models have been previously described but only at the level of planar models. In fact, the limited relevance of the presently known results for LW solutions has been pointed out by King, Owens and Vu [6] who have computed the contributions of the LW terms in the far field region. They found that a vertical dipole is superior to a horizontal one in exciting an ionospheric surface wave. However, the large directional gain that is achieved with the

long travelling-wave arrangement of horizontal dipoles in the wave antenna cannot be obtained with vertical dipoles.

Clearly in OTH radar systems, the curvature of the earth is of major importance. There is a standard starting point for this analysis due to Norton [16] who summarized most of the work that took place on the subject up until that time. An apparent exception is that of V. Fock [17], the Russian physicist, whose work later became available in the West [18]. This solution is exactly analogous to a Sommerfeld Integral Representation in the case of the "flat earth" approximation relevant to the Wu-King solutions that modify the radar equation as discussed above.

If LW contributions are to be important to OTH radar, it will most likely be in the "groundwave" mode. This is true because of the exponential damping with z that is characteristic of LW field components effectively rules out drastic modifications to the "skywave" mode. Still "groundwave" mode OTH is useful. Such a system operating in the 3-7 MHz range is capable of effective detection in the few hundred nautical mile range. Thus, "groundwave" mode OTH may be capable of filling one or more of the missions cited in a recent JASON report [18]:

The OTH applications that seem to us to be worth evaluating take advantage of the fact that OTH is one of only a few ground-based technologies having wide surveillance capability. Therefore it is important to evaluate the potential effectiveness of OTH against aircraft in a theater military application or in a counter-narcotics role; for the detection of ships ... a fleet defense role; for arms control monitoring and treaty verification; and for several interesting intelligence missions. For example, surveillance of Third World missile test ranges to determine operating characteristics, especially the existence and timing of possible submunition releases, would be a capability of high value in planning our next generation defense against theater ballistic missiles.

We note that this report was directed toward "skywave" OTH and it may be that there are inherent limitations on LW fields can be significant. To ascertain theoretically whether this is true is of some importance. Work has been done to model "groundwave" mode performance numerically (reported in [9]) using NEC-type capabilities. Still this subject to the "blindness" described earlier. Presently the theoretical characterization of OTH can be summarized following Skolnick [9] who has given a modified OTH radar equation,

$$\frac{S}{N} = P_0 \frac{G_a G_t T \sigma \lambda^2}{N_0 L (4\pi)^3 R^4} F_p .$$

here the propagation factor is clearly present. The factor that impedes such studies is that the Wu-King methodology has not been applied to the case of a conducting surface to our knowledge. Work is underway presently to fill in this lack of knowledge [20].

D. Implications for GPR Radar Studies

The role of the interface is crucial to the existence of LW field components. So clearly one wants to maximize the benefits of this effect. One other area where further work can be carried out is in the area of ground penetrating radar (GPR). This may be especially true in seeking unique signatures from objects buried not too deeply under the surface of the ground. In this arena it is the delicate interference pattern due to the interplay of the lateral wave with the direct wave that may offer new possibilities. Only now is a detailed study of integrating the phase history of an electromagnetic wave, including a LW contribution, underway [20].

E. Implications for Dipole-Sea Surface Studies

In a work completed last year [7], it was shown that there can be dramatic effects due to the inclusion of LW modifications. The explicit calculations of the the LW correction factors were for four situations. A radiating electric dipole (antenna) can be either vertically (i.e., along z-axis) or horizontally oriented. These two situations may be denoted by "ved" or "hed," respectively. Similarly, the polarization of the electromagnetic radiation may be either vertically or horizontally oriented to be respectively denoted by "v" or "h." So in total there are four *a priori* possibilities that are found to be:

$$\eta_{\text{ved/v}} = \left| \frac{1 + i n \frac{\lambda(h_a + h_t)}{2\pi h_a h_t} - \frac{(n^2+1)\lambda^2}{(2\pi)^2 h_a h_t}}{1 + i \frac{n^2(h_a + h_t)\lambda}{2\pi h_a h_t \sqrt{n^2-1}}} \right| ,$$

$$\eta_{\text{hed/h}} = \left| \frac{1 + i \frac{\lambda(h_a + h_t)}{2\pi h_a h_t} - \frac{\lambda^2}{(2\pi n)^2 h_a h_t}}{1 + i \frac{(h_a + h_t)\lambda}{2\pi h_a h_t \sqrt{n^2-1}}} \right| ,$$

$$\eta_{\text{hed/v}} = \left| \frac{h_t + i \frac{\lambda(n^2+1)}{2\pi n}}{n_a + h_t} \right| ,$$

$$\eta_{\text{ved/h}} = 0 ,$$

where n is the index of refraction for water or salt water, h_a and h_t are the respective heights of the antenna and target above the sea surface. The complex index of refraction n is a function of the frequency. Plots of these functions are given in reference seven, where it is shown that some of these can vary from 5–80 as the frequency of the signal goes from 30 MHz down to 3 MHz for the corrections associated with v_{ed}/v and v_{hd}/v in fresh water! Thus, the lateral wave contribution implies a huge anomalous enhancement at the low frequency (3–5 MHz) range. Other correction factors do not show such spectacular enhancement.

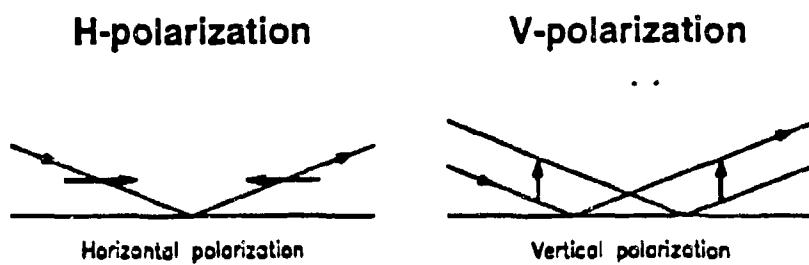


Figure 3. Diagrammatic Representation of Horizontal and Vertical Components of Dipole Radiation

6. Summary

We hope to have convinced the reader that the advent of the discovering of the Wu-King type of solution to the venerable problem of electromagnetic propagation near boundaries offers the possibility of new ways to achieve unexpectedly enhanced performance of systems that utilize these phenomenologies. There are a number of problems that only now are being addressed. Two such questions seem most pressing. Does the Wu-King method when applied to a spherical conducting earth give rise to an unambiguous lateral wave component just as it did in the flat earth approximation? In spite of the exponential decrease with height, is the lateral wave component robust enough so that the delicate interference patterns can be detected at some reasonable distance away? Both of these are purely technical questions now. When the answers to these questions are known, it should be possible to more realistically address the potential impact of the existence of lateral waves. This should be motivation enough to invest in finding the answers.

REFERENCES

1. J. Zenneck, "Propagation of Plane Electromagnetic Waves Along a Plane Conducting Surface and Its Bearing on the Theory of Transmission in Wireless Telegraphy," *Ann. d. Phys.* vol. 23, 846, 1907.
2. A. Sommerfield, "Propagation of Waves in Wireless Telegraphy," *Ann. d. Phys.* vol. 28, 665, (1909); idem. *Ann. d. Phys.* vol. 81, 1135, 1926.
3. T.T. Wu and R.W.P. King, *Radio Sci.* 17, 521, 1982.
4. T.T. Wu and R.W.P. King, *Radio Sci.* 17, 532, 1982.
5. R.W.P. King and M.F. Brown, "Lateral Electromagnetic Waves Along Plane Boundaries: A Summarizing Approach," *Proceedings of the IEEE*, Vol. 72, No. 5, 595, 1984.
6. R.W.P. King, M. Owens and T.T. Wu, *Lateral Electromagnetic Waves: Theory and Applications to Communications, Geophysical Exploration and Remote Sensing*, Springer-Verlag, New York, 1993.
7. G.N. Gilbert, E. Raiten and M. Visser, "Lateral Wave Contribution to the Low-Altitude Radar Propagation Factor," *Univ. of MD at College Park preprint*, UMDEPP 93-190, 1993.
8. F.E. Nathanson, *Radar Design Principles*, McGraw-Hill Book Co., NY, 1969.
9. M. Skolnick, *Radar Handbook*, 2nd Edition, McGraw-Hill Book Co., NY, 1990.
10. B. Edde, *Radar: Principles, Technology, Applications*, Prentice-Hall Inc., NY, 1993.
11. M.F. Brown, R.W.P. King and T.T. Wu, *J. Appl. Phys.*, Vol. 53, 3387, 1982; idem. *J. Appl. Phys.*, Vol. 55, 3927, 1984.
12. J.A. Kong, *Electromagnetic Wave Theory*, John Wiley & Sons, Inc., NY, 1986.
13. R.W. P. King and G.S. Smith, *Antennas in Matter*, MIT Press, Cambridge, 1981.
14. L.A. Berry and M.E. Chrisman, "A FORTRAN Program for Calculation of Ground Wave Propagation Over Homogeneous Spherical Earth for Dipole Antennas," *Nat. Bur. Stand. Rept.* 9178, 1966.
15. A.A. Kolosov et. al., *Over-the-Horizon Radar*, Artech House, Boston, 1987.
16. K.A. Norton, "The Calculations of Ground-wave Field Intensity Over a Finitely Conducting Earth," *Proc. IRE.*, Vol 29, 623, 1941.
17. V. Fock, "Diffraction of Radio Waves Around the Earth's Surface," *J. Phys. of the U.S.S.R.*, Vol. 9, No. 4, 255, 1945.
18. V.A. Fock, *Electromagnetic Diffraction and Propagation Problems*, Pergamon Press, Oxford, 1965, pp.191-212.

19. C. Callan, J. Cornwall, P. Diamond, S Drell, D. Eardley, S. Flatté, G. MacDonald, C. Max, F. Perkins, A. Peterson, J. Sullivan and J. Vesecky, *Advanced Over-the-Horizon Radar*, JASON Report No. JSR-90-105 (MITRE Corporation, McLean, VA, 1993).
20. G.N. Gilbert, private communication.
21. M. Abramowitz & I. Stegun, *Handbook of Mathematical Functions*, Dover Inc., NY, 1972.

Annex A

**EXPLICIT SAMPLE OF A SOMMERFELD INTEGRAL
REPRESENTATION**

Annex A
**EXPLICIT SAMPLE OF A SOMMERFELD INTEGRAL
 REPRESENTATION**

The explicit expression of the Sommerfeld integral representation of the component of the magnetic field of a unit dipole radiator for a vertical magnetic dipole on the boundary between two interfaces is;

$$H_{1z} = -i \frac{1}{8\pi} \int_{\text{SIP}} d\xi \frac{\xi^3}{\sqrt{k_1^2 - \xi^2}} \left(\frac{2\sqrt{k_1^2 - \xi^2}}{\sqrt{k_1^2 - \xi^2} + \sqrt{k_2^2 - \xi^2}} \right) H_0^{(1)}(\xi\rho) e^{i\xi z_1}$$

where $H_0^{(1)}(\xi)$ denotes the Hankel function of the first kind. This function has a standard representation in terms of a contour integral.

$$H_\nu^{(1)}(\xi) = \frac{1}{\pi} \int_{\Gamma_1} d\alpha e^{i(\xi \cos \alpha + \nu \alpha - \nu \pi/2)}$$

The contour Γ_1 can be found in many standard references (e.g., [21]).

Annex B
ORDINARY DIPOLE RADIATION COMPONENTS
IN FREE SPACE

Annex B

ORDINARY DIPOLE RADIATION COMPONENTS IN FREE SPACE

The unit electric dipole oriented along the z -axis has the following components for the electric and magnetic field as derived by standard methods.

$$E_{1\rho}(\rho, z) = -\frac{\omega\mu_0}{4\pi k^2} \frac{\rho z}{r^2} \left(\frac{ik^2}{r} - \frac{3k}{r^2} - \frac{3i}{r^3} \right) e^{ikr} ,$$

$$E_{1z}(\rho, z) = \frac{\omega\mu_0}{4\pi k^2} \left[\frac{ik^2}{r} - \frac{k}{r^2} - \frac{i}{r^3} - \frac{z^2}{r^2} \left(\frac{ik^2}{r} - \frac{3k}{r^2} - \frac{3i}{r^3} \right) \right] e^{ikr} ,$$

$$B_{1\phi}(\rho, z) = -\frac{\mu_0}{4\pi} \frac{\rho}{r} \left(\frac{ik}{r} - \frac{1}{r^2} \right) e^{ikr} ,$$

These relatively simple expressions should be compared to the analogous formulae in the case of the Wu-King solution given in the next annex.

Annex C
SAMPLE WU-KING DIPOLE SOLUTION COMPONENTS

Annex C

SAMPLE WU-KING DIPOLE SOLUTION COMPONENTS

The structure of the components of the electric and magnetic fields become significantly more complicated in the presence of boundaries. In the Wu-King solution it can be seen that there are explicit terms below that are proportional to $f(\rho; k_1, k_2)$ and $g(\rho; k_1, k_2)$ which have no analogs in the free-space solutions. The advantage of the Wu-King method is that these terms can be found with a minimum of effort. The traditional use of the Sommerfeld integral representation makes it much more difficult to obtain these terms. Below we give the field components for a z -directed electric dipole.

$$E_{1\rho}(\rho, z) = -\frac{\omega\mu_0}{2\pi k_1^2} \frac{k_2^2}{k_1} \left[f(\rho; k_1, k_2) e^{ik_2\rho} e^{ik_1(z+d)} - i e^{ik_1 r_2} \left(\frac{1}{\rho^2} + \frac{3i}{2k_1\rho^3} \right) \right]$$

$$-\frac{\omega\mu_0}{2\pi k_1^2} \left[\frac{e^{ik_1 r_1}}{2} \left(\frac{ik_1^2}{r_1} - \frac{3k_1}{r_1^2} - \frac{3i}{r_1^3} \right) \left(\frac{\rho}{r_1} \right) \left(\frac{z-d}{r_1} \right) \right]$$

$$+\frac{\omega\mu_0}{2\pi k_1^2} \left[\frac{e^{ik_1 r_2}}{2} \left(\frac{ik_1^2}{r_2} - \frac{3k_1}{r_2^2} - \frac{3i}{r_2^3} \right) \left(\frac{\rho}{r_2} \right) \left(\frac{z-d}{r_2} \right) \right]$$

$$E_{1z}(\rho, z) = \frac{\omega\mu_0}{2\pi k_1^2} \frac{k_2^2}{k_1^2} [k_2 g(\rho; k_1, k_2) e^{ik_2\rho} e^{ik_1(z+d)}]$$

$$-i \frac{\omega\mu_0}{2\pi k_1^2} \frac{k_2^2}{k_1^2} \left[e^{ik_1 r_2} \left(\frac{z+d}{\rho} \right) \left(\frac{ik_1^2}{\rho} - \frac{k_1}{2\rho^2} + \frac{7i}{8\rho^3} \right) \right]$$

$$+\frac{\omega\mu_0}{2\pi k_1^2} \frac{e^{ik_1 r_1}}{2} \left[\frac{ik_1^2}{r_1} - \frac{k_1}{r_1^2} - \frac{i}{r_1^3} - \left(\frac{z-d}{r_1} \right)^2 \left(\frac{ik_1^2}{r_1} - \frac{3k_1}{r_1^2} - \frac{i3}{r_1^3} \right) \right]$$

$$-\frac{\omega\mu_0}{2\pi k_1^2} \frac{e^{ik_1 r_2}}{2} \left[\frac{ik_1^2}{r_2} - \frac{k_1}{r_2^2} - \frac{i}{r_2^3} - \left(\frac{z+d}{r_2} \right)^2 \left(\frac{ik_1^2}{r_2} - \frac{3k_1}{r_2^2} - \frac{i3}{r_2^3} \right) \right]$$

$$B_{1\phi}(\rho, z) = -\frac{\mu_0}{2\pi} \frac{k_2^2}{k_1^2} \left[f(\rho; k_1, k_2) e^{ik_2\rho} e^{ik_1(z+d)} - i e^{ik_1 r_2} \left(\frac{z+d}{\rho} \right) \left(\frac{ik_1}{\rho} - \frac{3}{2\rho^2} \right) \right]$$

$$-\frac{\mu_0}{2\pi} \left[\frac{e^{ik_1 r_1}}{2} \left(\frac{ik_1}{r_1} - \frac{1}{r_1^2} \right) \left(\frac{\rho}{r_1} \right) - \frac{e^{ik_1 r_2}}{2} \left(\frac{ik_1}{r_2} - \frac{1}{r_2^2} \right) \left(\frac{\rho}{r_2} \right) \right]$$

$$f(\rho; k_1, k_2) = \frac{ik_2}{\rho} - \frac{1}{\rho^2} - \frac{k_2^3}{k_1} \left(\frac{\pi}{k_2 \rho} \right)^{1/2} e^{-ip} \mathcal{F}(p) ,$$

$$g(\rho; k_1, k_2) = \frac{ik_2}{\rho} - \frac{1}{\rho^2} - \frac{i}{k_2 \rho^3} - \frac{k_2^3}{k_1} \left(\frac{\pi}{k_2 \rho} \right)^{1/2} e^{-ip} \mathcal{F}(p) ,$$

$$r = \sqrt{\rho^2 + z^2} , \quad r_1 = \sqrt{\rho^2 + (z-d)^2} , \quad r_2 = \sqrt{\rho^2 + (z+d)^2}$$

$$p \equiv \frac{k_2^3 \rho}{2k_2^2}$$

$$\mathcal{F}(p) = \int_p^\infty \frac{e^{it}}{(2\pi t)^2} dt = \int_0^\infty \frac{e^{it}}{(2\pi t)^2} dt - \int_0^p \frac{\cos t + i \sin t}{(2\pi t)^2} dt$$

$$= \frac{1}{2}(1+i) - C_2(p) - iS_2(p) ,$$

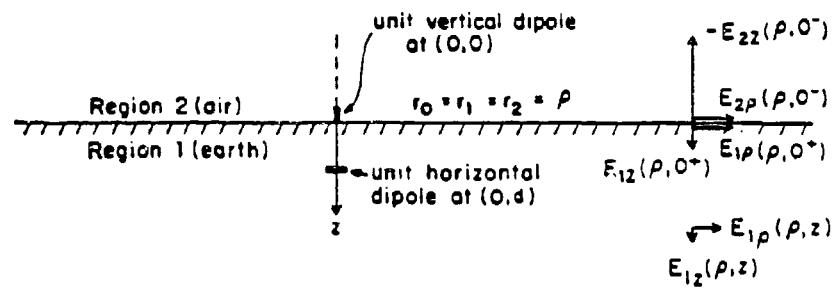


Figure 4. Definition of Electric Field Components for Unit Vertical and Horizontal Dipoles On and Below Air-Earth Interface

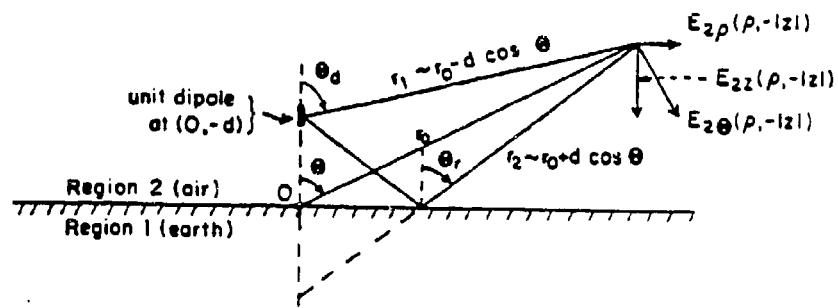


Figure 5. Definition of Electric Field Components for Unit Dipole Above Air-Earth Interface

LIST OF SYMBOLS

α	complex integration variable
$B_{1\phi}$	magnetic field azimuthal component in region-1
C_2	the second cosine Fresnel integral function
D	distance from source of EM radiation
d	height of dipole source above plane interface
E	generic component of electric field
E_0	component of electric field at source
$E_{1\rho}$	electric field radial component in region-1
E_{1z}	electric field z -component in region-1
η_{LW}	lateral wave relative correction factor
F	the complex Fresnel integral function
F^{total}	exact pattern propagation factor
F_{Fresnel}	Fresnel approximation of pattern propagation factor
$F_{a \rightarrow t}$	antenna-to-target pattern propagation factor
$F_{t \rightarrow a}$	target-to-antenna pattern propagation factor
f	a function of complex Fresnel integral
F_p	propagation-path factor
Γ_1	contour for first Hankel function

G_a	antenna gain
G_t	target gain
g	a function of complex Fresnel integral
H_z	magnetic induction field z -component
H_{1z}	magnetic induction field z -component in region-1
$H_0^{(1)}$	0-th order Hankel function of the 1-st kind
$H_\nu^{(1)}$	ν -th order Hankel function of the 1-st kind
h_a	height of antenna above sea surface
h_t	height of target above sea surface
i	the square root of -1
k	complex wave-number of signal
k	Boltzmann's constant
λ	wavelength of signal
L	transmission-path and system losses
μ_0	magnetic susceptibility of free space
N_0	noise power per hertz
n	complex index of refraction
$\Delta\omega$	a complex number classifying Hankel functions
p	2π times the frequency of signal
P_0	bandwidth (range) of frequencies in signal
P_r	Sommerfeld's "numerical distance"
R	radar transmitted power

ν	power of return signal
ω	distance from transmitter to target
r	spherical radial distance
r_1	first modified spherical radial distance
r_2	second modified spherical radial distance
ρ	cylindrical radial distance
$\rho_{LW\text{-}max}$	LW modified theoretical maximum range of detection
σ	radar cross section
S_2	the second sine Fresnel integral function
<i>SIP</i>	Sommerfeld integration path
S/N	signal-to-noise ratio
T	temperature
T	effective processing time
ξ	complex integration variable

**F. CONFLICT AND INTEREST:
SITING A NUCLEAR WASTE REPOSITORY**

Nancy M. Haegel
**Department of Materials Science and Engineering
University of California, Los Angeles
Los Angeles, California**
and
**Department of Physics
Fairfield University
Fairfield, Connecticut**

EXECUTIVE SUMMARY

Long term storage of high level nuclear waste has been a topic of study and debate since 1957. The United States, as a result of nuclear power plants and nuclear defense programs, has generated approximately 30,000 tons of high level nuclear waste that is now in temporary storage sites at reactors and government installations around the country. In 1987, the Department of Energy (DOE) was charged by Congress to characterize Yucca Mountain, a desert site in the state of Nevada, for suitability as a permanent buried geological repository. Significant opposition to the proposed facility exists in the state of Nevada and the siting process initially described by Congress in the Nuclear Waste Policy Act of 1982 has become increasingly politically and financially costly.

The technical and scientific advances that have allowed the country to utilize nuclear power for both commercial and military activities have not been matched by an ability to deal with the technical and social challenges presented by the production of radioactive waste products from these activities. Key public interests of health and safety, control, economic security and government credibility are at stake in the siting process. The Yucca Mountain Project can be viewed as a case study of how a government agency interacts with the public on complex technical issues with long term social, political and environmental implications.

Conflict resolution is an approach to dealing with contentious issues in which many and varied interests are at stake. The Harvard Negotiating Project, under Roger Fisher and William Ury, identifies four fundamental concepts of principled negotiation: 1) focus on interests, not positions; 2) invent options for mutual gain; 3) insist on using objective criteria; and 4) separate people from positions. A review of the history of the siting process for a nuclear repository shows that DOE and other parties in the process have followed a repeatable pattern of establishing programs and taking positions prior to serious consideration of the interests of potentially affected parties. This approach is guaranteed to produce maximum opposition and slow progress toward an equitable negotiated solution. The attempt to negotiate from positions rather than from interests is, I suggest, the primary flaw in the way that the nuclear waste repository issue has been approached for over 20 years.

Specific suggestions for action include 1) that DOE join other agencies funding scientific and technical work to create a network of scientific and technical professionals who are trained in and available for principled negotiation and 2) that the forms of public outreach and education being used today in the areas of nuclear waste, nuclear safety and the repository be adapted to deal more effectively and directly with the issue of risk. These actions are independent of the fate of the Yucca Mountain Project. If that project encounters further difficulty then the following actions may be required: 1) that citizen groups be asked to discuss the nuclear waste issue with an eye toward "inventing options for mutual gain," 2) that an NRC panel be convened which would go beyond the issues of fairness and equity raised in its recent reports and consider the process itself and 3) that more equitable processes be considered for a new repository siting program.

I. WHERE HAVE WE BEEN? 1945-1993

A. INTRODUCTION

The issue of nuclear waste storage and disposal is as old as the use of nuclear energy itself. Yet 36 years after the start of the first commercial reactor and, not coincidentally, 36 years since the National Academy of Sciences recommended a buried geologic repository for long term nuclear waste storage [1], no permanent facility exists in the United States for the storage of spent reactor fuel and high level radioactive waste.

Faced with a growing limitation on spent fuel storage at local facilities and a government commitment to begin accepting commercial nuclear waste for permanent storage or disposal in 1998, Congress turned its attention to this issue in 1982 with the passage of the Nuclear Waste Policy Act. This legislation designated underground repository as the primary U.S. approach for high level waste disposal and mandated the review of multiple sites across the country for such a facility. In 1987, this act was amended to prescribe that Yucca Mountain, a desert site 100 miles from Las Vegas, Nevada, was the sole site to be fully evaluated and characterized for a geologic repository. Site characterization activities are currently underway, and DOE estimates call for a repository, capable of isolating waste for 10,000 years, to be in operation in 2010, assuming that no reason is found to disqualify the site. Significant public opposition to the siting has arisen in the State of Nevada, and proponents and opponents of the plan continue to be engaged in a battle for public opinion.

In the 48 years since the dawn of the nuclear age, the U.S. has come to depend on over 100 nuclear power plants to generate approximately 22% of its electricity, and the defense establishment has routinely deployed, over the last 20 years, between 20,000 and 30,000 nuclear warheads. At the same time, the nuclear industry has generated 22,000 metric tons of spent fuel, most of which is in temporary storage at reactor sites [2]. Defense programs have produced another 8,000 metric tons of high level waste, stored in various levels of suitability at DOE sites around the country. The technical and scientific advances that have allowed the country to utilize nuclear power for both commercial and military activities have not been matched by an ability to deal with the safety, technical and

social challenges presented by the production of highly radioactive, long term waste products from these activities. The level of activity and progress toward reaching military and commercial goals contrasts greatly with the very limited progress in dealing with waste storage and clean-up issues.

Technological challenges, though abundant, are not the primary factor retarding progress on the issue of nuclear waste disposal. The following study will review the siting of the nuclear waste repository to understand the interaction between the public and governmental agencies on a public policy issue involving complex technical problems with long term environmental and social consequences. This will be done by reviewing the history of the siting process and then evaluating the process in view of the principles of conflict resolution, sometimes referred to as "principled negotiation." The primary goal of the study is to reassess the nuclear waste repository problem from a stance in which public interests are perceived as a starting point, rather than an obstacle, to a solution.

B . HISTORY

The history of the siting process for a nuclear repository is a long complex affair involving three agencies (the Atomic Energy Commission (AEC), its successor the Energy Research and Development Agency (ERDA) and its successor the Department of Energy (DOE)), the National Academy of Sciences (NAS), Congress and state and local officials. A brief summary of key events is presented in the chronology at the end of this section [3].

The first serious attempt to follow through on the 1957 NAS recommendation to evaluate salt formations as the optimum form of geologic repository was Operation Salt Vault which was begun in Lyons, Kansas, in 1963. The initial program involved temporarily burying seven canisters containing spent fuel to evaluate the effect of the heat and the radiation on the salt. No indications seemed to exist, either within the AEC or by the scientists involved in the project, that the Lyons site was being considered for a permanent repository. The technical studies proceeded quietly with the support of state officials.

Then in 1969 a fire at the Rocky Flats plutonium facility resulted in the production of 9,300 cubic meters of plutonium-contaminated waste, which was eventually shipped to the National Reactor Testing Station (NRTS) in Idaho for burial with other transuranic waste. When this information became public, Frank Church and other Idaho officials protested strongly and were told that all waste at the NRTS site would be removed and disposed of permanently in a federal repository. According to Luther Carter, author of an

influential 1987 survey of the nuclear waste problem, "Shortly afterward the AEC announced its tentative decision to convert the Lyons salt mine to a repository" [4].

The announcement produced a great deal of trouble and embarrassment for the AEC. The governor of Kansas was informed of the decision slightly before the announcement, but his approval or input on the decision was not requested. The proposal also undercut the work of a NAS panel studying the suitability of disposal in salt. The results from subsequent evaluations showed the Lyons site to be unsuitable for a variety of reasons and in September of 1971, work on the site ceased. The scientific issues which arose, mainly associated with a nearby mining operation and problems of water containment, affected the credibility of the AEC. The project was officially abandoned in 1973.

In summarizing this first repository effort and interaction with the public, Carter writes "The agency had underestimated the complexities and uncertainties inherent in repository siting, which had been made all the greater in this case by virtue of the extensive oil and gas exploration and salt mining in the region. It had also failed to establish from the start a cooperative relationship with Kansas state officials. The Lyons fiasco is still remembered as a testament to the technical hubris and political naiveté which the AEC brought to this first attempt at a delicate siting task" [5].

In 1973 another event occurred which set the stage for early relations with the public on the nuclear waste issue. Storage tanks at the Hanford facility were known to have begun leaking as early as 1958 and by 1973, 15 out of 149 single wall tanks were confirmed to be leaking into the surrounding soil. A new leak was discovered on June 8 in which 115,000 gallons of waste had escaped from the tank. The leak received national publicity and brought widespread attention to the issue.

Coupled with increasing public fear and suspicion was a growing concern about the viability of the commercial nuclear industry if the waste problem were not satisfactorily resolved. In 1976, California passed a law imposing a moratorium on new reactor production pending the development of demonstrated technology for the disposal of high level waste. This action raised the concerns of the nuclear power industry and publicly heightened attention on the waste disposal problem.

By this point, the ERDA was rethinking its waste disposal program. In 1976, the agency had sent letters to 36 governors informing them that it was planning to search for repository sites in their states and that 13 field investigations would begin before the end of 1977. A highly negative reaction was received. A DOE review of the initiative later

admitted that the plan was poorly designed with regard to state and local interaction and that "public concerns were aggravated rather than resolved" [6]. Within ERDA, the focus began to narrow to government-owned properties, with the Nevada Test Site and the Hanford reservation as prime candidates. One of the first tasks President Carter requested from the newly formed Department of Energy in 1978 was a review of its radioactive waste management program. The review concluded that significant research would be required before proceeding with a geologic repository and that the suggested 1985 operating date was not realistic.

Finally, beginning in 1979, the 96th and 97th Congresses turned their attention to the nuclear waste problem. DOE wanted a congressional mandate to proceed with the project, while the nuclear industry was also hopeful that Congressional intervention would move the project along. Similarly, the various potential host states were hopeful of being able to lobby to include provisions for consultation and ultimately veto power over the projects. The Nuclear Waste Policy Act (NWPA), which was finally passed in 1982, committed the nation to permanent geologic repositories for nuclear waste, with multiple sites to be evaluated. However, the Act also called for DOE to notify potential host states for the first repository with 90 days of the act's taking effect. This meant that only sites that had previously been considered, such as Hanford and Yucca Mountain, were likely to be included.

By 1986, DOE announced that three sites—Yucca Mountain in Nevada, Hanford in Washington and Deaf Smith in Texas—would be characterized for a first repository. Simultaneously, the search for a location for a second repository, focusing on sites in the Midwest, New England and the Southeast, was suspended. The reason given was that demand for electrical power was decreasing, so the development of a second repository could be delayed, but in fact major political opposition had arisen in all these areas of the country, prompted in part by a site selection process that had allowed sites within close proximity to major population areas and drinking water sources to appear on the initial lists.

The feeling that their selection has been preordained all along, coupled with DOE's abandonment of the search for a second repository site, caused significant reaction in the West, and all three states that had been selected as potential sites filed suit against DOE soon after the sites were announced. The process appeared headed for both political stalemate and ballooning characterization costs. In 1987, Congress amended the NWPA to mandate that only Yucca Mountain would be fully characterized as a potential site for the repository.

The state of Nevada passed a law opposing the siting of the repository and refused to provide permits required by DOE to proceed with characterization activities. In 1990, DOE sued the state in federal court and won. In 1993 work was begun on what will ultimately be a 5-mile tunnel that is a key part of the geological evaluation. Annual appropriation for the project is approximately \$375 million.

The goal of the program is to build a repository that can accommodate up to 70,000 metric tons of radioactive material. Regulations for the design are based upon EPA standards that require radioactive nuclide release limits resulting in less than 1,000 deaths among 10 billion people over 10,000 years. Current plans call for characterization activities to proceed from 1993-2000. If, at that point the site is deemed suitable, licensing and construction will proceed. The most optimistic date for the opening of the repository is 2010. Should the site be deemed scientifically unsuitable, for any reason, the DOE must go back to Congress for further instruction.

So stands the process which the AEC set in motion almost 40 years ago.

Table 1. Chronology

Year	Activity
1955	AEC requests National Academy of Science (NAS) study on waste disposal
1957	NAS recommends geologic disposal; Start of commercial nuclear power
1963	Project Salt Vault begins studying suitability and geology of salt formations
1970	AEC policy for commercial high level waste proposed including federal geological repository and federal responsibility for ultimate disposal
1970	Lyons, Kansas, salt mine announced as first repository
1973	Lyons, Kansas, abandoned as potential site in view of technical, political problems
1973	Major radioactive leak at Hanford—15 out of 149 tanks known to be leaking
1976	California imposes moratorium on new reactors pending development of demonstrated technology for disposal of high level waste
1976	Letters sent to 36 governors informing them that sites in their states will be considered for a repository
1978	DOE reviews waste management program and concludes that significant additional research is required before proceeding with a repository program
1982	Nuclear Waste Policy Act passed by Congress legislating underground disposal and requiring characterization of multiple sites
1985	Three sites selected by DOE for further characterization: Deaf Smith, Texas; Hanford and Yucca Mountain
1987	Nuclear Waste Policy Amendment Act; only Yucca Mountain for further study
1989	Nevada passes law opposing the repository. Refuses to process permits for DOE
1990	DOE sues the state of Nevada over permit issue and wins
1993	Tunnel excavation begun at Yucca Mountain for characterization
1993–2000	Proposed site characterization period
1998	Target date for government to receive spent fuel from nuclear power industry
2001–2009	Expected licensing and construction period
2010	Earliest possible date for opening of the repository

II. WHERE ARE WE NOW? CURRENT KEY ISSUES AT YUCCA MOUNTAIN

Before moving into the areas of public interest, public trust and the role of conflict negotiation principles, we will summarize briefly the primary subjects of current debate in the nuclear repository program.

A. CURRENT SCIENTIFIC/TECHNICAL ISSUES

1. The Water Table and Water Migration

Current plans call for the repository to be situated 1,000 feet below the surface, in the "unsaturated zone" above the water table. One of the positive features of the Yucca Mountain site is the depth of the water table, which is currently an additional 800 feet below the proposed repository site. Controversy arose when DOE geologist Jerry Szymanski claimed in an internal memo to have found evidence from mineral deposits that the water table had once risen near to the surface and, therefore, could do so again within the 10,000 year timeframe of interest [7]. Subsequent studies by independent groups, including a NAS panel, have not supported this claim, but it remains a point of contention with those who oppose further site characterization efforts. The movement of water down from the surface is also a major issue, because ground water is the most likely means of both corrosion of the engineered barriers for the radioactive waste and the transport of released radionuclides.

2. Earthquake and Volcanic Activity

There are reportedly 32 known active faults in the Yucca Mountain area and some earthquake activity could be expected. DOE states that experience shows that underground structures are highly stable with respect to ground motion and that the primary earthquake safety concerns would involve the surface facilities. The agency claims that current assessments indicate the earth movement has been minimal and that studies show the region has remained "largely unchanged over the last million years" [8].

Yucca Mountain itself is the product of volcanic activity. Today there are seven small dormant volcanoes in the area, two of which may have been active within the last 10,000 years, according to DOE information. Critics raise the question of the likelihood of volcanic activity which could affect the integrity of the repository, while DOE suggests that the probability of an eruption is "very remote" and, in its public literature, places the chances of volcanic activity directly affecting a repository at "about 1 in 500 million per year" [8].

3. Engineered Versus Natural Barriers

The isolation of nuclear waste at a repository will depend upon both engineered or manmade and natural barriers. Current regulations call for the waste packaging to contain the spent fuel during the first 300 to 1,000 years in the repository. Beyond that point, it is expected that some materials degradation will occur and that natural barriers to radionuclide movement, including interaction with the rock and the relative dryness of the region, will be important to limit the movement of the radioactive materials.

Sweden is also pursuing a buried repository program, though nothing is expected to be operational before the year 2020. Sweden's nuclear program is limited, though the country now depends on nuclear power for roughly 45% of its electricity. The Swedish repository design, however, calls for greater dependence on engineered rather than natural barriers and suggests placing the spent fuel in thick Cu canisters designed for much longer integrity than those proposed for the U.S. program. Many groups, including the National Research Council (NRC), have criticized the heavy reliance of the U.S. program on natural barriers.

B. SOCIAL/POLITICAL ISSUES

1. Yucca Mountain as Sole Site for Characterization

A primary source of concern for both the future of the repository program in general and the integrity of the Yucca Mountain project in particular is the 1987 Amendment to the Nuclear Waste Policy Act designating Yucca Mountain as the sole site to be characterized for a potential first repository. The basis for concern is obvious. Should the Yucca Mountain site be found unsuitable, for any reason, DOE must return to Congress for further instruction. Since no other site is being investigated in parallel, this would set the process back to the beginning and delay the opening of any repository for an extended period of time.

The decision to concentrate solely on Yucca Mountain had a variety of motivations. Because of its deep water table, its presence on federal land and remoteness from developed areas and the expected support of the local populace, Yucca Mountain probably was the most suitable of the three sites being considered for complete characterization activities. An extensive study of the waste management problem, *Nuclear Imperatives and Public Trust*, published in 1987, came to this same conclusion [9]. However, the small size of the Nevada congressional delegation, coupled with the powerful positions of Thomas Foley of Washington and Jim Wright of Texas, led many to believe that Yucca Mountain was also politically most expedient, since its legislators would be unable to stop the bill. This led to widespread resentment in the state of Nevada, and the 1987 amendment is still referred to as the "Screw Nevada Bill."

Public concern is that the absence of any backup plan may bias DOE to find the Yucca Mountain site suitable and to slant its results toward that end. DOE responds that the high degree of oversight on the project (from the state of Nevada, the Nuclear Regulatory Commission, the EPA, the Nuclear Waste Technical Review Board and the National Academy of Sciences) precludes this. As a matter of process, however, this issue affects the public's perceived level of trust in the project.

2. The Role of Public Opinion

Opposition to the repository siting tends to be greater in the northern part of the state than the southern [10]. Communities near the site, currently dependent in part upon jobs at the Nevada Test Site, are hopeful that the repository would offer the prospect of new jobs. Having lived with the test site, many residents also feel better educated about the hazards of radiation and are more likely to believe that a repository can be located near their communities with minimal risk. The Las Vegas area, with its heavy dependence on tourism, has remained largely uncommitted during the discussion, with some people suggesting that if the presence of the test site has not deterred visitors, it is unlikely that a repository would do so either.

A widely publicized campaign has been occurring in the state in recent years in an attempt to form public opinion. On the one hand, anti-nuclear groups, some Nevada papers and state and national officials who are opposed to the siting present articles and talks opposing the project and raising public awareness of possible safety issues and DOE mismanagement at other facilities. On the other side, the nuclear industry, represented by the American Nuclear Energy Council (ANEC), and DOE attempt to counter what they

view as unfair and unsupported negative publicity. The ANEC has supported a campaign of television commercials, seminar and lobbying activities, and DOE maintains three Yucca Mountain information centers for the public which offer educational displays on nuclear power, radiation, and the repository.

C. CONCERNS OF BOTH A TECHNICAL/SCIENTIFIC AND SOCIAL/POLITICAL NATURE

1. DOE Credibility and Public Trust

Responsibility for a large-scale long-term program which is highly dependent on public sentiment and trust could probably not come at a worse time for the DOE. Growing public awareness of the environmental problems at DOE facilities around the country has reduced public trust in the DOE's ability and commitment to safely manage radioactive waste. Some estimates for the cost of DOE clean up efforts go as high as 200 billion dollars. Billions of gallons of radioactive waste were inappropriately disposed of at the Hanford reservation and most scientists acknowledge that the clean-up efforts there present a massive challenge, in part because of ignorance of what was done in the early years. In Nevada, evidence exists that AEC officials may have been aware of the danger to military and civilian personnel from atmospheric nuclear testing between 1951 and 1963 [11].

Both DOE and the National Research Council have admitted that the credibility of the agency with the public is very low and a major issue in the repository program. Efforts to include public input and oversight in the Yucca Mountain project are one response, but it is unclear that claims of a "fresh start" are sufficient to deal with the accumulated distrust of recent years. From both a technical as well as political standpoint, however, the leaking tanks at Hanford and the plutonium contamination at Rocky Flats stand as ever-present participants in the Yucca Mountain discussions, undercutting DOE assertions of adequate knowledge and/or commitment to guarantee long-term waste isolation.

2. Excessive and Unrealistic Regulations and Requirements

One response to public concerns about the program has been a high level of regulation and review for the Yucca Mountain Project. This has led to a highly detailed set of program requirements in place long before significant characterization research was performed. According to a 1993 article in the *Dallas Morning News*, over 27,000 requirements exist, in one form or another, for the project [12]. Overseers of the program, including the National Research Council panel, suggest that more flexibility is required to

acknowledge and allow for uncertainties that are bound to exist when engineering for such a long timeframe. The NRC report (1990) said

The U.S. program is unique among those of all nations in its rigid schedule, in its insistence on defining in advance the technical requirements for every part of the multibarrier system.... In this sense, the government's high level waste program and its regulation may be a "scientific trap" for DOE and the U.S. public alike, encouraging the public to expect absolute certainty about the safety of the repository for 10,000 years and encouraging DOE program managers to pretend that they can provide it [13].

The question of the extent and validity of our knowledge and predictions over such long timeframes is an important one in the public debate.

III. WHAT DO WE WANT? PUBLIC TRUST AND PUBLIC WILL

Before introducing the principles of conflict resolution and their application to the repository siting issue, it is important to discuss some of the fundamental interests and concerns that motivate public opinion on the topic. These can be summarized under the general headings of 1) health and safety, 2) control, 3) economic security and 4) trust and credibility. These reflect many of the hierarchy of needs identified long ago by Abraham Maslow [14].

Because identifying and meeting such needs is an integral part of the successful implementation of any solution, it is a mistake to underestimate their importance in the forming of public policy. This is a mistake which scientist and engineers, by nature of their disciplines and training, are likely to make. People involved in conflict resolution often find it necessary to remind such professionals that "emotional" data is valid data when it comes to the formation of public policy.

The following sections summarize the various human interests at stake in the siting process.

A. HEALTH AND SAFETY

This is the primary interest which generally receives the most publicity and discussion in the siting process. This is a multi-level concern, involving not only individuals but also their sense of responsibility for the health of future generations. The lack of definitive studies on low level radiation risks, coupled with the public's inability to monitor their own exposure, makes this a highly problematic area of debate.

In the view of many scientists, the public has an "irrational" fear of radioactivity. One point often mentioned is that the public is also not familiar with the idea of risk and risk assessment, leading to unreasonable demands for absolute safety and health assurances. This misses the point. Although the public may not fully understand the methods used to generate some of the probabilities used in decision making, they are actually quite familiar with the idea of risk assessment. Performing hazardous work, having a baby, flying in an airplane, and choosing elective surgery are all choices people make. They give their consent, in light of significant benefits, knowing that the event is

outside their complete control (or that of anyone else) and that some level of risk is involved. This type of comparison—rather than those that are more commonly made to things like the background radiation level, accidents in the home, and driving—are probably more appropriate and more useful in public education and debate. After all, most people have to be on the earth, get to work, and live a house. They do not have to welcome a nuclear waste repository into their community.

B. CONTROL

One of the most basic of human needs is to exercise control over one's environment. The issue of control is especially important in the repository siting issue for two reasons. First, because radioactivity cannot be seen or routinely monitored by the average person, people are much more likely to feel a lack of control than they do about other hazards which they can see, feel, smell, etc. Second, because the radioactive material will be placed at a site by the federal government, all choice is removed from individuals, except to move from the area, with regard to exposing themselves and their children from whatever level of hazard may exist. That is why the common comparison of the health risks associated with living near the repository with activities such as driving or flying is not fully persuasive, since such comparisons do not take into account the issues of real or perceived control and perceived benefit.

C. ECONOMIC BENEFIT

In most cases, one takes on voluntary risks because of perceived benefit. This benefit could be social, economic, or personal. What is important initially is that the perceived benefit significantly outweigh the perceived risk. Over time, real benefit and real risk can be more easily evaluated.

The question of economic benefit associated with a repository is a very important one. In addition to jobs associated with characterization, construction, and operation of the repository, there is the issue of direct compensation to the state and its people for, in a sense, performing a service to the nation. The perceived economic benefit may vary widely through the state, differing for example between the local community and more distant population centers. This is the general case for the Yucca Mountain project, where local community familiarity with the nuclear testing program and dependence on DOE as an employer have a long history.

In discussing the question of additional economic benefit in exchange for the repository, there are two issues that must be acknowledged. Many critics view the discussion of

economic benefits as bribery and are fundamentally opposed on principle. However, one must also realize that there is no direct benefit to a majority of the people of the state in hosting the repository, and so it is difficult to imagine why they would choose to accept any level of imposed risk, however minimal, without some economic or other benefit.

D. TRUST

Tempered by the reality of experience and human nature, people want to have a level of trust both in the public officials they elect to speak for them and the experts they entrust with various tasks. In the case of the repository, the credibility of both DOE and various scientists and engineers are major factors in the discussion. A major series of articles on the Yucca Mountain Project, published October 21-28, 1990 in the *Las Vegas Review-Journal* was entitled: "Yucca Mountain: A Question of Credibility" [15].

Various factors come into play here. First and foremost is the issue of DOE's track record in dealing with the public and the growing public awareness of environmental problems, of varying magnitude, at DOE weapons facility sites across the country. Although discussion of this appears nowhere in public relations literature concerning Yucca Mountain, it is an issue DOE cannot ignore. Also important is the changing perception of science and scientists in recent years. Luther Carter, reflecting upon the early days of the nuclear age, writes "The Atomic Energy Commission and the 'atomic scientists' carried a mystique, and the public viewed them with awe. While this was not the same thing as trust, for a time it would be a serviceable substitute" [16].

The atmosphere of secrecy and the lack of public debate that characterized the beginning of nuclear energy and nuclear power meant that the industry began in an era of public ignorance about nuclear technology. Clearly both the perception of the scientist and the level of concern about nuclear safety have changed. Even assuming complete DOE openness and impeccable behavior with regard to the public trust, the level of trust and credibility, slowly eroded over a 40-year time period, cannot be suddenly regained.

These issues of public interest are common to many policy issues. Similar concerns would arise with regard to the siting of any hazardous chemical plant, any disposal facility, prisons, etc. The nuclear waste disposal issue is unique however because of its potential to affect generations far beyond our own, further than we have come this far in the annals of written history. In addition to the commonly identified "stakeholders" in the process, future generations are present as unspoken participants in the process, required to deal with the waste in whatever form they receive it.

IV. HOW SHOULD WE PROCEED? PRINCIPLES OF CONFLICT NEGOTIATION

In this section, we will summarize principles of conflict negotiation as a framework for reassessing the issues involved in the siting of a nuclear waste repository. The reference for this part will be the work of the Harvard Negotiating Project, as summarized by Fisher and Ury in their 1981 book, *Getting to Yes* [17]. Their approach has been used, in various forms, in government, business, and labor and international forums, as well as individual and community negotiations.

A. FOCUS ON INTERESTS, NOT POSITIONS

The most important concept in principled negotiation, as opposed to what is often called standard positional bargaining, is that principled negotiation focuses first on the interests of all parties involved. Standard positional bargaining often begins with set, and generally conflicting, positions. Fisher and Ury illustrate this important distinction with the following story:

Consider the story of two men quarreling in a library. One wants the window open and the other wants it closed. They bicker back and forth about how much to leave it open: a crack, halfway, three quarters of the way. No solution satisfies them both.

Enter the librarian. She asks one why he wants the window open: "To get some fresh air." She asks the other one why he wants it closed: "To avoid the draft." After thinking a minute, she opens wide a window in the next room, bringing in fresh air without a draft [18].

This story is typical of many negotiations. Since the parties' problem appears to be a conflict of positions, and since their goal is to agree on a position, they naturally tend to think and talk about positions—and in the process often reach an impasse. The librarian could not have invented the solution she did if she had focused only on the two men's stated positions of wanting the window open or closed. Instead she looked to their underlying interests of fresh air and no draft.

In most negotiations, there will be multiple interests held by all parties. Fisher and Ury point out that the most powerful issues are issues of basic need, including security,

economic well being and control over one's life, interests which we have identified as major actors in the nuclear waste repository siting issue.

In the case of the repository siting, some key interests are clear. DOE is interested in having a safe, long-term facility that can aid it in its dual challenges of security concerns for nuclear nonproliferation and its commitment to accept the spent fuel from the industrial sector for long-term disposal. The nuclear power industry is interested in seeing DOE fulfill its commitment (thereby relieving the industry of a major problem) and in removing this obstacle to the future of nuclear power. Environmentalists are interested in preserving the local environment and assuring the health and safety of the population. The general public, both locally and more broadly, have a variety of interests which have been previously discussed, foremost among them the basic needs of security, control and economic well being. Note that having or not having a repository *at Yucca Mountain* is a position, one which may meet some needs while threatening others. It is not, in and of itself, a fundamental interest for anyone.

B. INVENT OPTIONS FOR MUTUAL GAIN

The primary point here is to arrange negotiations which tend to broaden, rather than restrict options. A variety of options exist for doing this, including brainstorming, identifying shared interests, focusing on successful precedent, etc. An example of inventing options for mutual gain involves the negotiations between Egypt and Israel over control of the Sinai Peninsula. The negotiated agreement—a demilitarized Sinai controlled by Egypt—was achieved by broadening the options, generally envisioned as either one side or the other maintaining full militarized control of the land. The new option met the Israeli need for security and the Egyptian need for right to its ancestral land [19].

Primary obstacles which Fisher and Ury identify which inhibit the production of options for mutual gain include 1) premature judgment, 2) searching for a single answer, 3) the scarcity assumption (assumption of a "fixed pie") and 4) the tendency to focus solely on one's own position and interests and leave the other party to do the same.

C. INSIST ON USING OBJECTIVE CRITERIA

The reality of conflict negotiation is that interests do conflict. When this occurs, it is suggested that some agreement on external or objective criteria be used in place of a battle of will and power among the participants. In the case of business negotiations this objective criteria is often something like market value or the result of arbitration.

In the case of Yucca Mountain, criteria for issues, such as the reliance on engineered versus natural barriers, how many health effects are reasonable, what kind of compensation is acceptable, would require the participation of independent organizations. The National Academy of Sciences has played this role for a long time in the scientific arena, and the EPA has a role to play. Similar organizations must be found for the issues of economic compensation which may arise.

D. SEPARATE PEOPLE FROM POSITIONS

In all negotiations, no matter how bureaucratic, it is important to remember that we deal with human beings, not, as the authors remind us, with "abstract representatives of the 'other side'." Both the substance of the final agreement and the relationship are important if the groups are to work together in the future. Recognizing the importance of the people also emphasizes the role of emotion. Under principled negotiation, emotion should be explicitly recognized and acknowledged as legitimate.

In the case of Yucca Mountain, the legitimacy given to the emotions of fear and distrust which affect public opinion is very important. This is an emotional issue, not a purely technical one. It affects how people feel about their safety and that of their children. At the same times, scientists feel under attack for what they consider objective work. Political emotions in Nevada run high. It is clear that an isolated technical solution is neither possible nor desirable for such an issue.

The guidelines for principled negotiation are, at some level, obvious. Yet our patterns of behavior and negotiation—as individuals, organizations and nations—indicate that we often forget them in practice, choosing a more common strategy of positional bargaining, supported by acts of will, strength, and ultimately force. Having reviewed these principles, we turn now to an assessment of the nuclear waste repository siting issue in view of the model.

V. HOW HAVE WE DONE? A CRITIQUE OF CURRENT POLICY

In the siting of a permanent high level waste repository, DOE has been charged with a task unprecedented in its technical and social challenges. The time scale of the project and the need for public accountability make this a task unlike any other that any government agency has undertaken. And although interaction with the public has been an issue since the beginning of the nuclear power industry, DOE operates in a significantly different environment and with a much higher degree of oversight than was the case in the 1970s. For DOE, the interaction with the public required for its civilian waste management program represents a major difference from the nuclear programs it has operated for the Department of Defense. DOE has responded to many of the concerns and demands associated with the project by holding open public hearings, sponsoring stakeholder meetings, performing public education and being attentive to the many organizations involved in oversight functions for the Yucca Mountain project. Yet much of this activity is fairly recent, and to evaluate the repository issue, one must consider a longer history that affects public perception and trust.

The most striking fact when one reviews the repository issue is that there has been a repeated history of violating the first and most basic of the conflict resolution principles. From the Lyons salt mine project through initial site selection attempts in the 1980s and continuing through the Yucca Mountain project, the AEC, DOE and, most recently, Congress have adopted a pattern of "decide, announce, defend." The pattern has been to announce a position or plan of action prior to extensive public consultation, and then to attempt to involve communities and obtain the local or state support required to proceed. With regard to the initial repository program at Lyons, then with the three initial sites under the NWPA, then with the case of a second repository and still potentially with Yucca Mountain, there is a repeatable pattern of an announced position or program, followed by significant local or regional opposition, followed by a DOE reassessment or retreat.

The principles of conflict resolution tell us that this is a path guaranteed to produce maximum opposition and to slow progress toward an equitable negotiated program. This is because by announcing a plan of action (even if only a tentative or preliminary one) prior

to extensive consultation with local communities with regard to implementation, the responsible agent has begun by taking a position, i.e., *we want to consider a repository here*. The response to this is for the other parties to take the opposite position, i.e., *we don't want it here*. Conflict and opposition are created by the initial announcement, positions are staked out and stalemate is quickly reached—all before any discussion of actual interests has occurred.

To illustrate this in more detail, consider both an early example—the Lyons, Kansas, case—and a more recent example—the 1987 amendment to the NWPA. Scientific progress was occurring in the Operation Salt Vault program. There was little opposition and a reasonable degree of support existed among state officials and local scientists. Scientists at Oak Ridge had consulted with and considered the interest in involvement of the local community. The AEC's sudden announcement that the Lyons site would be considered for a permanent repository broke the process of consultation and created strong opposition. The project, as we have seen, was ultimately abandoned for technical reasons, but the political climate and AEC credibility required for proceeding were seriously jeopardized. Had the AEC initiated local discussion on the question of a permanent site prior to its announcement, they would have learned of the mining history of the site and other concerns which ultimately stopped the project. By moving to a position, prior to discussing the interests of state and local communities, they destroyed the basis for further negotiation and needlessly damaged their credibility in the process.

Similarly, 17 years later, Congress directed DOE to follow a similar path when it amended the NWPA to designate only Yucca Mountain for site characterization. No widespread consultation, no agreement on acceptable risk and benefit, no state involvement in the selection process occurred prior to the Amendment. Procedures for all these things were put in place once the site characterization was to begin but, at that point, the political and social damage had already been done. Positions were taken and a full discussion of interests was thereby abandoned, resulting in the complex tangle of regulation, distrust and delay that exists today.

The attempt to negotiate from positions, rather than from interests is, I believe, the primary flaw in the way that the nuclear waste repository issue has been approached for well over 20 years.

Other problems in the process become apparent when viewed from the principles of conflict resolution. By designating only Yucca Mountain to be considered, the Congress has caused DOE to focus on a sole solution, limiting the range of options available for

public discussion. The decision to focus on an underground repository was primarily a technical decision made by a relatively small group of people and then presented to the public. With a topic of such social and political import, a wider group involved in the process at an earlier stage would have led to a wider consensus, even if the final outcome were unaltered.

Finally, there has been limited attempt, by groups on either side, to give the public objective criteria to use in making decisions on this topic. As the NRC study pointed out, "it is essential to bear in mind that the comparison is not so much between ideal systems and imperfect reality as it is between a geological repository and at-surface storage" [20]. Yet this issue is rarely discussed, and the NRC panel goes on to suggest that the EPA, which played a key role in setting standards for the project, "has not based its standards on social judgments derived from realistic considerations of these alternatives" [21].

A number of positive steps can also be identified which are incorporating the insights gained from conflict resolution. DOE maintains an extensive program of statewide meetings, stakeholder conferences, and public hearings. The use of independent review teams, usually under the auspices of the NRC, have played a key role along the way. And the NRC, in its 1990 report, included in its deliberations a category entitled "Moral and Value Issues" in which they recognized the "central role of a fair process." The NRC panel looked at questions concerning the professional responsibilities of scientists and engineers on these types of issues and the use and limitations of science in the decision making process. The panel stopped short, however, of critiquing the process itself.

VI. WHERE DO WE GO? LESSONS FOR TOMORROW

Given its difficult history and contentious present, the siting of a nuclear waste repository at Yucca Mountain may be one area where "remedial" conflict resolution cannot work. As discussed in Chapter IV, the ideal of negotiating from interests to positions, as opposed to the other way around, has been violated so repeatedly on all sides in the history of the siting process that the trust required to remedy the situation may no longer exist among the parties.

Yet many scientists increasingly believe that on-site storage is a relatively safe short-term (i.e., 50–100 years) option. If this is true, then the siting of a permanent repository may be an important test case of governmental willingness to proceed in a different fashion in the future with regard to the development of nuclear facilities. Stopping the Yucca Mountain Project in favor of a new process for decision-making and siting would have large financial and political costs, while the inventory of high level waste continues to grow. On the other hand, the history and current situation suggest that stalemate may occur anyway, while a unique opportunity to bring to bear all our knowledge of problem solving (as opposed to solely our technical knowledge and political savvy) may have been lost. With this issue, the Congress and DOE have a chance to set a new direction for the country on the larger issue of interaction with the public on complex technical issues with long-term social, political and environmental implications. One can only think that this type of problem will increasingly present itself in the next millennium.

Suggestions for specific action include:

1. That DOE join with DOD, NSF, NIH and other government agencies funding scientific and technical work to create and support a network of scientific professionals who are trained and willing to participate in conflict resolution—These people could be identified and trained with grants from various agencies and then be available to serve in this capacity for local communities, states and national government agencies. Such an initiative would illustrate national recognition, as evidenced in the NRC report on high-level radioactive waste disposal, that "exploration of ethical issues can illuminate the fundamental policy debates" on a variety of technical policy issues.

2. That the forms of public outreach being used today to educate the public on nuclear waste, nuclear safety and the need for and nature of a repository be adapted to deal more directly with the issue of risk—More appropriate analogies need to be made which differentiate between voluntary and involuntary risk. The public, as the NRC report stated, views the risk associated with nuclear waste as potentially "catastrophic, uncertain and involuntary." These attitudes require respect and an admission that the latter two may well be correct perceptions. One useful tool for education about radioactivity and its isolation is the concept of "natural analogs," the ways by which concentrated radioactive materials that already exist in the environment are "stored" by nature.
3. That an NRC panel be convened which would go beyond the issues of fairness and equity raised in its recent reports and consider the siting process itself—This panel could be composed of both experienced scientists and experts in conflict resolution.
4. That citizen groups be asked to discuss the nuclear waste disposal situation with an eye toward "inventing options toward mutual gain"—Groups should be asked to consider this question in every state in the nation, thereby including those with current storage facilities, those benefiting from the nuclear power, those with potential repository sites and those far removed from the current debate. This collection of voices may yield new insights on the concerns that must be addressed before a successful siting process can occur.
5. That a "reverse Dutch auction" and other more equitable processes be considered for a new repository siting process—There has been discussion of the use of financial compensation in siting "undesirable facilities (also called "locally unwanted land use" or LULUs) [22]. One proposal is the reverse Dutch auction in which the value of the compensation is increased until a state (or some other defined entity) accepts the compensation bid to site the facility. This is similar to the process which is used to deal with the overbooking of seats on airplane flights. The value of providing incentive in this fashion is that it allows the state accepting the facility to weigh its own alternatives and set its own price. It also means that the community is more likely to cooperate in the process. Suitability studies would be done by external bodies, with no compromise of standards, so states are more likely to bring forth sites that are believed most suitable.

One objection to this approach is that it may cause those states or areas with the greatest financial need to become the most likely sites for unwanted facilities, a regressive form of distribution of the nation's burdens. This aspect of the siting of LULUs goes on now, with or without the aspect of direct financial compensation. Although this approach is less than ideal, for this reason, it

does assure that no community will have the repository forced upon them if they themselves do not choose to consider the option. This protection is more than currently exists for many sparsely populated and politically less powerful regions of the country.

The reverse Dutch auction concept requires public discussion and input, but it could be a starting point for discussion, beginning the process of "inventing options for mutual gain" which is one of the key tenets of principled negotiation.

REFERENCES

1. National Academy of Sciences, "The Disposal of Radioactive Waste on Land," NAS-NRC publ. 519 (Washington, D. C.: National Academy of Sciences) Sept. 1957.
2. U.S. Department of Energy, "DOE's Yucca Mountain Studies," DOE/RW-0345P, December 1992.
3. Luther J. Carter, *Nuclear Imperatives and Public Trust* (Washington D.C.: Resources for the Future, Inc.) 1987. This is the primary source for the history survey.
4. Ibid., p. 67.
5. Ibid., p. 70.
6. U.S. Department of Energy, "Report of the Task Force for Review of Nuclear Waste Management," 1978.
7. William J. Broad, "A Mountain of Trouble," *New York Times Magazine*, Nov. 18, 1990.
8. U.S. Department of Energy, "DOE's Yucca Mountain Studies," p. 9.
9. Carter, p. 422-427.
10. S. Papinchak and L. Wingard, "Nuke View Changing," *Las Vegas Review-Journal*, October 21, 1990, p. 1.
11. L. Wingard and J. A. Morrison, "Confidence in DOE Low Despite Watkins' Vows," *Las Vegas Review-Journal*, October 24, 1990, p. 1.
12. D.J. Swanson, "Cost, Frustrations Soar as Nuclear Project Lags," *The Dallas Morning News*, May 23, 1993, p. 1.
13. National Research Council, "Rethinking High-Level Radioactive Waste Disposal," (Washington D. C.: National Academy Press) 1990, p. 1.
14. Abraham H. Maslow, *Motivation and Personality* (New York: Harper and Row) 1970.
15. Papinchak and Wingard, et. al., October 21-28, 1990.
16. Carter, p. 44.
17. Roger Fisher and William Ury, *Getting to Yes: Negotiating Agreement Without Giving In* (New York: Penguin Books) 1981.
18. Ibid., p. 41.
19. Ibid., p. 42.
20. National Research Council, p. 9.
21. Ibid., p. 15.
22. Herbert Inhaver, "How to Slay the Nimby Dragon," Proceedings of the 26th Inter-society Energy Conversion Engineering Conference (LaGrange Park, IL: American Nuclear Society) 1991.

**G. NEW DETECTOR SYSTEM FOR AIRPORT SECURITY
AGAINST BOMBING THREATS**

**Peter Chen
Harvard University
Cambridge, Massachusetts**

**Mark E. Davis
California Institute of Technology
Pasadena, California**

NEW DETECTOR SYSTEM FOR AIRPORT SECURITY AGAINST BOMBING THREATS

A. INTRODUCTION

What is the threat?

With the end of the Cold War and the disintegration of the Soviet Union and Warsaw Pact, some threats to U.S. national security have eased, but others may increase. Representative of the latter are terrorist acts against either military or civilian targets. The threat addressed in this initial study is the terrorist bombing of civilian aircraft and airports. Although there are a number of civilian-sector entrants in the airport security arena, a fresh examination of the technical constraints placed on the terrorist by the nature of the target warrants another look at possible ways to detect and defeat a bombing with minimum loss of life or property.

Fortunately for those who design strategies for the detection of high explosives, there are a relatively small number of chemical structures, manufactured by known routes, that are suitable for use in terrorist applications. The constraints of high energy-to-weight, low sensitivity to shock, stability to a wide range of environmental conditions, and availability, all contrive to limit the number of target compounds for a detection scheme. The explosive formulations most likely to be employed are the plastic explosives and dynamites, with the former being the bomb material of choice. Two common plastic explosive formulations are C-4, produced and used by the United States and allied armies, and Semtex, produced in the former Czechoslovakia and widely distributed in the Warsaw Pact countries for a range of legitimate applications. Over 960 tons of Semtex were sold to Libya in the '70s and '80s and subsequently distributed to radical groups worldwide. The bomb that downed PanAm 103 over Lockerbie with the death of all aboard was believed to a few (one or two) pounds of Semtex [1]. Among the 21 tons of high explosives sold to Libya by former CIA agent Edwin Wilson in 1977 may have been RDX-based formulations such as United States C-4 military demolition charges. These have also been distributed to terrorist organizations. A bomb made from between one-quarter and one-half pound of either C-4 or Semtex nearly downed TWA 840 over Athens, the mistimed

detonation being the major reason for the survival of the aircraft despite the deaths of four passengers sucked out through the hole in the fuselage. Other formulations, such as DuPont DataSheet, come as flexible sheets that can be carried onboard (instead of being checked in with luggage) in an ordinary manila envelope, and can effectively destroy the aircraft in flight with substantially less than one pound of explosives if placed next to an exterior wall.

The strategies described in this report will be aimed at C-4, Semtex, and commercial dynamites because of the likelihood that these will be the bomb materials used in most scenarios, and also because any other plausible bomb would be made from formulations having enough in common with one of these that they would be detected as well. The need for detection schemes will continue even though taggants are being introduced into explosives, including current production runs of Semtex, because the large inventory of high explosives available on the world market will continue to provide materials for those who wish to avoid making easily detectable bombs. Another class of bombs, ammonium nitrate/fuel oil (or ANFO) bombs, are not considered in this report because, although they may be suitable for targets such as the World Trade Center, they are not small enough or powerful enough to have utility in an aircraft application.

What are the design requirements or constraints?

An overall concept for an effective airport explosive detection system will involve screening of checked baggage, carry-on baggage and passengers [2]. The solution to this security problem will require a "systems" approach that will use a combination of different technologies that do not allow terrorists to know which detectors a particular threat will encounter [3]. By combining various modes of detection, the strengths and weaknesses of particular technologies can hopefully be compensated by one another in the system.

The three types of screenings each presents particular problems for the detection of explosive materials. For checked bags, the Federal Aviation Administration (FAA) requires a minimum rate of 600 bags/hour (6 seconds/bag) for explosives detection systems [2]. However, the airlines desire a much greater rate of screening (a typical Boeing 747 takes on 700–800 pieces of checked luggage [2]). Carry-on baggage and passenger screening for metallic materials is currently accomplished using X-ray-based and magnetic-based detection systems, respectively. We are all acutely aware of the time constraints imposed by these types of detection systems. Additionally, exposure of humans to high energy X-rays or neutrons is not acceptable. Thus, an airport explosive detection system

must address all of these issues in a cost effective manner that imposes little inconvenience for passengers.

Passenger inconvenience is a critical issue since the number of threats in a particular year is extremely small relative to the total passenger traffic. The number of checked bags transported in a particular year has been quoted to be on the order of 10^9 while the number of threats is of the order of 10. These numbers also show the magnitude of the problem; the vast majority of baggage pose no threat. The false-alarm rate of a detection system is normally correlated to the sensitivity. Reducing thresholds to detect smaller quantities of explosives most likely will result in an increased false-alarm rate—typically false positive responses (detector signals the existence of a threat that is not real). However, if the threshold is too high, false negative responses will occur (detector fails to alarm on a real threat). Thus, an explosive detection system must minimize false negatives while keeping the number of false positives to a level that does not impede passenger traffic. The "systems" approach to this problem is a good compromise. That is, a high number of false positives are acceptable at a first-level of detection. These items could then be analyzed by subsequent levels of detection that become increasingly more selective and costly since as the number of items decreases, the cost, sensitivity, and detection time can all increase.

One could envision a security system as follows. For passengers, a non-evasive detection system will be necessary, e.g., vapor detection. Multiple detectors could be used for carry-on baggage in a manner similar to the X-ray systems currently in use, e.g., add dual-wave X-ray backscattering to the current X-ray detectors. For checked baggage, inexpensive, slow detection systems could be installed at check-in counters. High false alarm rates that are false positive are acceptable at this stage. Next, the baggage that fails the initial screening could then be subjected to additional, more expensive, detection schemes with slower throughputs (shared devices), e.g., X-ray devices and thermal neutron activation (TNA) systems. Additionally, baggage received from individuals possessing certain characteristics (subjective decision at the check-in counter) could be placed through the complete screening procedure. One of the key issues is to develop an inexpensive, highly sensitive detection system that has a reasonable false-positive alarm rate for use at check-in counters. It would be helpful if this device could be modified for use as hand-held detectors for passenger screening with the appropriate adjustments for false-alarm rates. In this report, we address this issue.

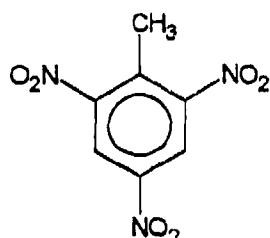
B. TECHNICAL BACKGROUND

What are the principal explosives to be detected?

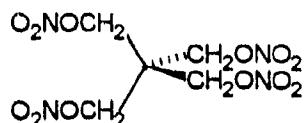
Formulations from various national sources use the same basic explosives. Only a few are used in nearly all military or commercial high explosives. The components in several formulations, including those mentioned in the Introduction, are tabulated below. Structures for the abbreviations, e.g., RDX, EGDN, are given later in this section.

Formulation	Main Explosive Components	Other Explosive Components (e.g., plasticizers, impurities)	Inert Components
C-4	RDX 91%	—	di-(2-ethylhexyl) sebacate 5.3%, polyisobutylene 2.1%, motor oil 1.6%
Semtex	RDX 44.5%, PETN 44.5%	EGDN	n-alkanes (C14 to C18) 10%
DataSheet	PETN 63%	nitrocellulose 8%	elastomer binder 29%
Cyclotol	RDX ~70%, TNT ~30% (typical)	DNT	hydrocarbon waxes up to 4%
Dynamite	nitroglycerin, ammonium nitrate, (proportions vary)	EGDN	wood pulp, kieselguhr

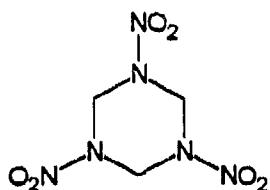
The structures of the constituent chemicals, along with CAS registry number and other names, are shown below. The original syntheses for all of the constituent compounds discussed here date back to the Second World War. There have been some improvements in the large-scale production of these explosives, but all countries use similar routes, which makes the detection problem easier.



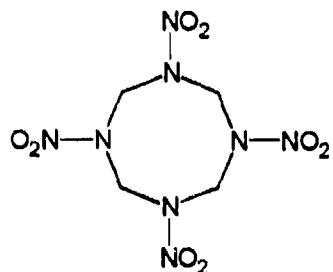
TNT (118-96-7)
2,4,6-trinitrotoluene



PETN (78-11-5)
pentaerythritol tetranitrate



RDX (121-82-4)
cyclotrimethylenetrinitramine
cyclonite
hexogen



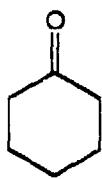
HMX (2691-41-0)
cyclotetramethylenetrinitramine
octogen

RDX is the most important constituent of plastic explosives. HMX is used in higher temperature specialty applications and is comparatively less common. Each of these is a target for detection, but a perusal of their physical and chemical properties quickly makes clear the difficulties that will be encountered in that task, especially if a vapor detection scheme is attempted.

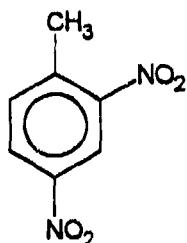
What are the targets for a vapor detection scheme?

While the compounds above are the principal components in most high explosive formulations, a number of minor constituents with considerably greater volatility are also present as additives, plasticizers, sensitizers, or simple impurities. These are also targets for a detection scheme, and, in fact, are easier to detect in many cases than the main explosive constituent. For example, United States military grade TNT contains up to 0.5% DNT from incomplete nitration [4]. Because of its much greater volatility, it is the principal vapor that outgasses from TNT samples. Rats trained to detect TNT by odor have been shown [4] to be reacting, in fact, to DNT. One compound, cyclohexanone, is not an explosive, but rather is a common impurity in nearly all RDX and some HMX. In the modern, large-scale preparation of RDX, removal of impurities that compromise

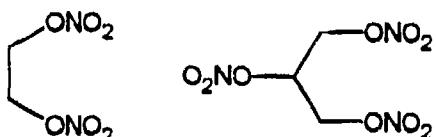
performance and storage qualities is done by recrystallization from cyclohexanone [5]. The recrystallization also affords control of the particle size distribution [6] in the final explosive, which is an important process parameter, and decreases shock sensitivity in the final product [7]. Accordingly, RDX contains up to 0.2% cyclohexanone in the form of occluded solvent within the RDX crystals. Photomicrographs of recrystallized RDX in an index-matching fluid (to improve contrast) reveals internal void spaces [8] in the crystallites that are filled with the cyclohexanone solvent. Drying does not remove the occluded solvent, but the final processing step of densification by mechanical pressing with a desensitizer/binder (usually a hydrocarbon wax or polymer) does fracture the crystallites [7] and release the cyclohexanone which can then outgas from the final product. Trials at Fort Belvoir have determined that cyclohexanone could be detected in the air around assembled mines using (what are now considered) relatively insensitive analytical techniques. Because of its high volatility, it is an obvious choice for vapor detection, although interference from non-bomb sources does not make it an unambiguous marker.



(108-94-1)
cyclohexanone
6 Torr @ 25°C



DNT (121-14-2)
2,4-dinitrotoluene
0.07 mTorr @ 25°C



EGDN (628-96-6)
ethylene glycol dinitrate
70 mTorr @ 25°C

NG (55-63-0)
nitroglycerin
1.8 mTorr @ 25°C

EGDN is a relatively volatile nitrate ester that is often used as an explosive plasticizer or a sensitizer in some explosives formulations. It is a common additive in commercial dynamites. Claims that RDX in Semtex had been detected by vapor methods have been shown to most likely be attributable to detection of EGDN instead [9]. EGDN has no non-bomb applications and hence should be a good marker. Nitroglycerin is used in commercial dynamites either as the main explosive constituent or as one of the major components.

Because it is highly unlikely that any terrorist group would be willing or able to make custom explosive formulations, at least one of the volatile additives or impurities ought to be present in just about any bomb. The additives and impurities are therefore the target of vapor detectors. One limitation on the detection schemes for these components is that nitroaromatics and nitrate esters are found in polluted urban air in the form of low-

concentration vapors, aerosols, and small particulate matter. Any detection scheme for explosives containing EGDN or DNT, for example, would have to distinguish between those and peroxyacetyl nitrate (PAN) and other atmospheric nitrates.

If the main explosives have such low vapor pressure, why do "sniffers" work at all?

Because of the low vapor pressure of the main explosives like TNT or RDX, the means by which they are detected is not their vapor, but rather dust on which the explosives are adsorbed. The use of a particulate collector/concentrator/detector as a front-end component of a bomb detection device is a novel aspect of this report. The non-imaging-type bomb detectors have relied on vapor detection of the explosives themselves, which a perusal of the vapor pressure data indicates to be wildly unrealistic. The transport of detectable traces of an explosive is undoubtedly by way of small mechanically produced particles of the explosive or by way of physisorbed vapors on the surface of dust grains. The successful detection of bombs by vapor "sniffers" (like the EG1S) is probably due either to the detection of more volatile secondary components, e.g., EGDN, or the adventitious detection of dust-borne explosives that are entrained in the air entering the device. A device designed from the start to optimize the collection, concentration, and detection of dust-borne explosives traces has significant advantages in sensitivity, selectivity, and, as will be seen later, in the ability to circumvent countermeasures.

How are dust-borne explosives distributed and transported?

In a CIA study [4] of the transport of explosives vapors and dust, an initially clean room with well-defined air flow patterns was set up. A small sample of military grade TNT was placed in the room. Sampling of the air found that TNT concentrations plateaued at a level on the order of 10 pg/L within 2–3 days. After 1 year, the air concentrations were not significantly higher. After removal of the TNT sample, the TNT level in the air remained high, taking 100 days to drop by a factor of three. Interestingly, every surface in the room was found to be contaminated with TNT, with each 4 cm² section of the wall (or floor or ceiling, all covered with latex paint) loaded with 10–200 ng TNT after 1 year. An approximate mass balance indicated that essentially all of the "evaporated" TNT resided on the walls of the room. After withdrawal of the TNT sample, the concentration of TNT on the walls declined with a 6–12 month half-life. The sampling methodology used in the CIA study did not distinguish between *bona fide* TNT vapors, TNT particles, and TNT adsorbed on other dust particles—any TNT on particles would have been detected if the

particles were suspended in the sampled air. The best rationalization for the CIA results is that by far the greatest fraction of the TNT was adsorbed on surfaces, with the very high surface area dust (typical surface area [10] of ~25,000 cm²/g for suspended particulates at 100–200 µg/m³ in outdoor urban air [11]) being the primary mode by which material is transported site-to-site. A similar study using radioactive TNT in a sealed chamber found that the vapor never reached equilibrium because of continual surface adsorption. In that study, localization of the radioactive tags found the explosives distributed over all surfaces in the chamber. As long as most of the dust is on a surface at any given time, the airborne concentration of TNT will be low. Importantly, even after a bomb is packaged or removed from the room in which it was assembled, explosive-contaminated dust particles on the walls of that room will serve as a reservoir to contaminate anything which stays in that room for even as short a time as a few days. In a striking example cited in the CIA study, rats trained to detect TNT (actually DNT) by smell could also detect RDX because the RDX had been stored in the same room as a TNT sample, resulting in the contamination of the RDX by minute amounts of DNT-containing TNT.

Why is dust an ideal target for detection of explosives?

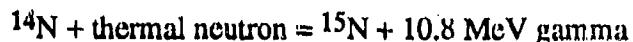
The size distribution [12] of dust particles has three maxima at particle aerodynamic radii of 0.1, 0.8, and 15.0 µm, corresponding to condensation aerosols, ash particles, and coagulated particles, respectively. Only the largest of these particles settle out by gravitation. The smaller particles are much more numerous than the larger ones, and contribute most of the surface area on which adsorbed material can be transported. The smallest particles also have a near-unity sticking coefficient when they impact upon a surface. Once they are on the surface, they may be removed mechanically, i.e., by being dislodged by another object, or be "blown" off and re-entrained into the air. In a room with a sample of TNT, one can envisage the small dust particles impacting on the explosive and becoming contaminated. Physical handling of the explosive, as well as air currents (for the larger particles), re-entrain the dust particles which eventually find their way to the walls where they stick. Collisions with other dust particles transfer some of the adsorbed TNT to those other particles and distributes the contamination over the whole room.

A very important feature of particles adhering to a surface is that they are difficult to re-entrain into the air without a mechanical disturbance of the surface. Air flow over a dust-covered surface of a suitcase, for example, under any normal conditions of transportation will not remove the dust. A wind tunnel study [13] found that only 21% of fly ash and 5% of atmospheric dust particles with a radius of 0.7 µm on a smooth surface

could be re-entrained into air with a bulk flow velocities up to 150 m/sec. Under conditions of laminar air flow over a surface, there is a stationary boundry layer in the air at the surface, and farther out, a parabolic velocity (vs. distance from the surface) profile up to the bulk velocity. Small particles on the surface are in the boundry layer and hence feel no force from the air flow. At high flow velocities (Reynolds' number >10,000) or for very rough surfaces, the boundry layer is disrupted and eddies at the surface re-entrain particles, but no such flow conditions occur in the transport of a suitcase. The suitcase, in which a bomb has been hidden, will be contaminated on the inside and the outside, by virtue of having been in the room where the bomb was assembled. Any mechanical process involving cutting, rubbing, or bending of the explosives will put an even heavier loading of explosives onto dust, which will then find its way onto the exterior of anything or anyone in the room. It will not matter that the bomb has been carefully wrapped in plastic to prevent vapors from escaping as the evidence of its presence will be all over anything that has been in its proximity.

What kinds of explosives detectors are currently available?

Three general types of detection systems are currently being tested in various airports throughout the world: nuclear-based devices, X-ray-based devices and vapor/particle detection devices. Of the nuclear-based devices, TNA has received the greatest attention. The basic idea with TNA is to use Californium-252 to generate thermal neutrons that react with ^{14}N to give off gamma rays:



The gamma rays are then detected. Since the active components of explosive materials have high nitrogen contents, TNA is capable of detecting low amounts of explosive components. However, TNA detects any material containing ^{14}N so selectivity is always an issue [2]. A 4-month test [14] of TNA detection was conducted at JFK airport in 1989. The TNA system used was able to investigate 600 bags/hour and was tested for 173 days scanning 52,413 bags. The alarm rate was 12%; however, the system was operating in a mode to trigger on the "slightest doubt" about a given item. When used in combination with a second method, e.g., X-ray-based system or a second pass through the TNA device, the alarm rate dropped to 2%. This alarm rate is well within the standards set by the FAA (~5%). The cost of a TNA device is approximately \$1 million. Thus, TNA may be used as a component of the overall security system but is more likely to be shared by numerous airlines.

Another nuclear-based technique that may show promise in the future is pulsed fast neutron activation (PFNA) [15]. A PFNA device would cost around \$2 million but would reduce the false alarm rate to near zero and could interrogate entire Boeing 747 baggage carts at one time. PFNA uses a nuclear particle accelerator to generate fast neutron radiation to provide three-dimensional spatial resolution of all the elements within the interrogated area. Thus, PFNA could provide increased probability of detection with simultaneous reduction in false alarms.

Numerous X-ray-based devices exist. For example, EG and G Astrophysics Research Corp. sells an X-ray transmission system for explosives detection [16]. The device distinguishes between anything of atomic number above and below 10 (explosives have a lot of N and O). An operator must identify the explosive by a color image on a computer screen. Because of inherent problems with transmission X-ray detection, American Science and Engineering has developed an X-ray backscattering device [16, 17, 18]. The detection system uses a combination of transmission and backscatter images (on two screens). Since low atomic number materials produce more backscatter than high atomic number solids, the backscatter image highlights low atomic number regions of the bag. In a research trial, the system was able to automatically recognize 1 kg bombs 100% of the time with a 2% false alarm rate [17]. Additionally, this detection system had an 80% success rate of identification for 0.32 kg bombs with a similar false alarm rate. This type of system can scan approximately 3,600 bags per hour which is well above the FAA standard. Tests of transmission and backscattering detection systems at airports in Zurich, Glasgow and other cities in the U.K. are currently underway [19]. These devices vary in price but are typically in the hundreds of thousands of dollars range. Thus, X-ray-based detection systems look very promising for checked and carry-on baggage. For checked baggage, an X-ray-based system used in combination with TNA would provide excellent detection capabilities. With carry-on bags, the combination of transmission and back-scattered X-rays appears very promising.

Vapor detection systems can be the only real answer for screening of passengers. This avenue for explosives admission to aircraft must not be overlooked [2]. Additionally, this type of system has the potential of being low cost. The problem, of course, is the low volatility of explosive materials. Ion Track Instruments, Inc., markets a hand-held vapor detector [16] that costs around \$40,000. The system uses two gas chromatography analyses (one to monitor the explosive vapor, the other to eliminate it) in order to identify vapors from explosives. It is surprising that this system is being explored, since the vapor

pressures of explosive materials are very low and the primary mechanism of airborne transmission of these compounds is more likely from adsorption on particulate matter [20]. In this regard, Thermedics, Inc., has developed a vapor/particle detection system [16, 19, 20] called EGIS. The hand-held units cost around \$150,000 while the complete devices can be as high as \$500,000. EGIS is able to detect the presence of plastic explosives vapor in concentrations as low as 0.01 ppt. The EGIS hand-held unit collects air at 2 liters per second and the sample is exposed to a projector lamp to heat the surfaces of the particulate matter to 67 °C. This temperature is sufficient to desorb components such as RDX, HMX and PETN. These vapor-phase species are then subjected to a series of chemical tests including separation by gas chromatography and analysis by the chemiluminescent reaction of NO (from pyrolyzed organic nitrates and nitro compounds) with O₃. Total analysis time is about 30 seconds. Thermedics also manufactures a walk-in unit for passenger screening. Warm air is blown over the passenger in a closed compartment and the exhaust air is analyzed. Currently, a person must stand in the device for 5–10 seconds and the analysis time is 30 seconds. Newer versions of this system are aiming for screening times of 10 passengers per minute. All of the Thermedics systems have been field tested or are currently being field tested in airports throughout the world. For example, a test was conducted in October 1988 at Boston's Logan airport. The test involved 2,000 passengers and only one false alarm was observed and was found to be caused by a computer problem [21].

C. PROPOSED DESIGN

We proposed that a new vapor/particle detection system be constructed following the general design presented below. This system is designed to be a low cost, highly sensitive detector for use as a distributed, front-end unit. The device should reveal a minimal number of false negatives but is allowed to have a high false positive rate since it is expected to be used as the "first line of defense" in an overall airport security system.

The new vapor/particle detection system involves two subsystems. The first subsystem is used to collect and concentrate airborne particulate matter into a localized area (spot). The spot will be exposed to a heat lamp to desorb any low-volatility components, e.g., HMX, RDX, PETN, that will be transported (pulse of components in a clean air stream) into the second subsystem for analysis of the vapor-phase components. The second subsystem utilized an array of sensors to detect the presence of vapor-phase compound used in plastic explosives—both the active components and residual materials. The vapor-phase separated from the particulate matter in the first subsystem is analyzed as

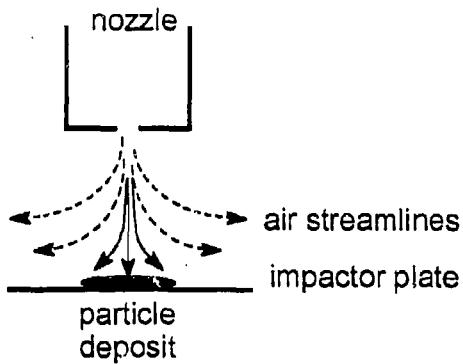
well as the desorbed material from particles collected in subsystem one. Below we describe the two subsystems in details.

Subsystem 1: Sampling strategies for dust

Because the 0.1 to 1.0 μm dust particles that are contaminated with explosives adhere to the surface of whatever was in the room with the bomb and are not removed by ordinary air flows associated with transportation and handling, efficient collection of a dust sample for analysis is an important design target. For hand-screening of specific sites on a person or object, a simple wiping of the surface with a slightly roughened nylon pad, as used by the FBI for narcotics detection on a surface [22], would collect whatever dust was on that surface. While effective, the procedure is poorly suited to automation and may not be acceptable to the public. An alternative to removal of the dust by mechanical wiping is re-entrainment into air by a high-velocity jet directed at a surface. As long as the stagnation pressure of the jet was above \sim 2000 Torr (absolute), the jet would be supersonic, and upon impact with a surface, should create turbulent eddies that would re-entrain any dust. Operation in a pulsed mode would give the same effect with lower overall gas load. The resulting air, with suspended dust, can then be exhausted into a detector system under conditions of laminar flow to keep the dust from contacting any surfaces along the way. This approach is well-suited to automation and can still be used with spatial resolution of a few centimeters. The suspended dust from a large volume of air will then be concentrated and collected for analysis with a cascade particle impactor.

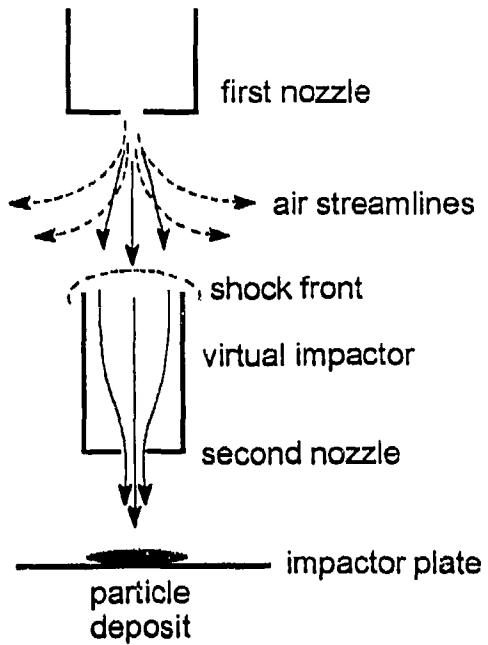
Subsystem 1: Particle Impactors: Basis for operation

An inertial impactor [23] is an aerodynamic device for separating suspended particulate matter from air. Suspended particles that are too small to be efficiently removed by sedimentation (settling under gravity) follow the streamlines of the air flow. Moving air exerts a force on the particle that depends on that particle's aerodynamic size. For low velocities, viscous forces dominate and the particles follow streamlines in the flow even as the streamlines curve around an object placed in the flow. At a higher velocity, the momentum of the particle takes it across streamlines if they curve. The simplest inertial impactor consists of a subsonic nozzle from which air and suspended particles flow, and a flat plate that is positioned directly in front of the nozzle. The air streamlines curve around the plate, but particles with a mass/diameter ratio above a certain cutoff cannot make the curve and impact onto the plate. If the plate is coated with grease, a visible deposit may be seen after collection has proceeded for a time.



The impactor has the effect of taking the particles that were in a large volume of gas, and collecting them at a single spot.

A refinement of the concept called a virtual impactor [24, 25] allows a construction of cascade impactors that have well-defined upper and lower bounds on the particle sizes collected. In this variation, the real impactor plate is replaced by a shock front normal to the flow. A well-defined shock surface can be created if the nozzle is operated as a supersonic, rather than subsonic, nozzle and a blunt tube is placed in the flow. The gas streamlines curve away from the mouth of the tube, creating an effect identical to that in the case of a real impactor plate. In the virtual impactor, though, the particles above the cutoff size, punch through the shock surface and enter the tube, which can be coupled in turn to another nozzle and virtual impactor stage. In this fashion, a cascade, or multistage impactor [26], can be constructed to pass only the desired size cut. At each stage of a virtual impactor, the particles above the cutoff size are transmitted but most of the air is diverted. Multiple stages accordingly give a flow that is highly enriched in suspended particles. A final, real impactor plate then serves to collect the particles that were suspended in a large volume of air onto a very small spot.



If explosives are carried by dust particles, then a cascade impactor serves to concentrate and collect those dust particles to maximize sensitivity. Moreover, because most of the air is diverted away at each stage, the cascade impactor will be sensitive only to particulate matter and not to vapors that may be in the flow. Properly designed, the selectivity for a certain size cut can be used as an additional selection criterion if it can be shown that one particular size dust particle is especially good for explosives transport. This added measure of size selectivity may also help discriminate against particles of condensed nitroaromatics that occur in some polluted urban environments.

The design and operation of both real and virtual impactors, singly and in cascades, is well-documented in process control and air quality management applications [27]. Commercial units that give >99% collection efficiency collection of particles in a size cut from 0.1 to 1.0 μm are available. For use in the collection of explosive-contaminated dust, the large volumetric flow rate needed to run a cascade virtual impactor is an advantage in that it translates into the concentration of the dust from a very large volume of air onto a spot in a short period of time.

Subsystem 1: Post-deposition detection of the explosive

Once the particulate matter contaminated with explosives is collected onto a spot by a cascade particle impactor, the detection and identification of the explosive components can be achieved by thermal desorption of the explosives residues by heating to ~100 °C.

Because of the high surface area and low thermal mass of the dust particles, the desorption can be done quickly to form a highly enriched pulse of gas for detection. For the purpose of economy in design and operation, the detectors can be the same polymer-coated SAW devices used in the vapor-phase detector module (see next section) of the proposed design.

Subsystem 2: SAW chemical sensors for vapor detection

The vapor detection system utilizes an array of surface acoustic wave (SAW) chemical sensors. Each sensor will serve a particular function that when integrated together will provide information regarding the presence of HMX, RDX, PETN, EGDN, DNT and cyclohexanone (solvent occluded in HMX, RDX, C-3, C-4, SEMTEX). The reason for this choice of sensor array is that the SAW devices are cheap, sensitive, and have very fast response times. The problem with these devices is that they are not chemically selective. Thus, to add selectivity to the SAW sensor, thin films of specially designed polymers will be coated on the SAW devices to provide the molecular recognition function.

Piezoelectric crystals have been used as sensors for quite some time. However, it wasn't until 1979 that Wohltjen and Dessimme measured *chemical vapors* with coated SAW crystals. The primary reason for using SAW devices rather than bulk resonating crystals (like in wrist watches) is that the frequencies can be much higher in the SAW device and the wave is localized in the superficial region of the crystal. Since adsorption of vapor-phase species occurs on the surface, the SAW devices are more sensitive than bulk devices at the same frequency, and the higher the frequency, the larger the sensitivity. Thus, SAW devices are much more sensitive than bulk wave devices. For example, Bowers *et al.* report a 200 MHz SAW device [28] that has better than 1 Hz frequency stability and mass sensitivity of 9×10^{10} Hz cm²/g. Additionally, Bowers *et al.* were able to detect HCl vapors in the ppb range using a SAW device coated with a triethanolamine film. Thus, SAW devices are small size (~cm²), low cost, low power consumption, fast response, highly sensitive detectors for monitoring gas phase species. As mentioned previously, the major drawback of these devices is their selectivity. The selectivity must be imparted by the addition of a thin film.

Subsystem 2: Imparting chemical selectivity to the SAW device

Numerous reports have appeared concerning polymer-coated SAW devices as selective sensors. In general, the selectivity is limited to classes of compounds rather than a specific target molecule. For example, an array of four SAW sensors, each having

polymers with different properties (hydrophilic, hydrophobic, acid, base), has been developed at NRL to detect chemical warfare agents. One sensor is used to measure humidity since water normally affects the properties of most polymers. The remaining SAWs respond differently to compounds containing different functional groups, e.g., alcohol, acid, amine. Thus, a "fingerprint" response to a variety of species is obtainable with this system. However, it is doubtful that the system could do a good job in distinguishing a CW agent from insecticides. Here we will use the concept of templating polymers to achieve higher selectivity in the films to be placed on the SAW devices.

The SAW array must contain a sensor for monitoring humidity since water normally affects the physicochemical properties of polymers. This is routinely done by a polyethyleneimine-coated SAW device.

The compounds to be detected are HMX, RDX, PETN, DNT, EGDN and cyclohexanone. The vapor pressure of these compounds at ambient conditions is given below:

Compound	Approximate Vapor Pressure @ RT and 1 atm.
HMX	0.1 ppt (0.1 in 10^{12})
RDX	1 ppt
PETN	1 ppt
DNT	0.5-0.1 ppm (0.1 in 10^6)
EGDN	100-10 ppm
cyclohexanone	20,000 ppm

Assuming a dilution factor of 10^{-3} (amount of saturated vapor to the total amount of collected vapor), DNT, EGDN and cyclohexanone should be detectable by a SAW device (can detect ~ppb) without any pre-concentration. Obviously, HMX, RDX and PETN are well below the detection limits without concentration. For HMX, RDX and PETN, the first subsystem will collect and concentrate particulate matter that would contain these compounds and then desorb them in a concentrated pulse. This concentrated pulse will bring the amount of these species to a detectable level. Currently the EGIS system desorbs these components from particulate matter and then sends the vapor through a gas chromatograph for separation. Here it will not be necessary to separate these species. The strategy is to develop selective sensors to recognize individual species. Thus, selectivity is the key to this design.

The concept for obtaining selectivity is to develop polymer films that have been templated to recognize certain species. The concept is nicely illustrated by the work of Mosbach *et al.* [29]. Functional monomers (methacrylic acids, MAA) are mixed with a "print" molecule (theophylline in the Mosbach report) and cross-linking monomers (ethylene glycol dimethacrylate) and the system is polymerized. The MAA is chosen for its ability to form hydrogen bonds with a variety of functional groups in "print" molecules. After the polymer is formed, the "print" molecules are removed by extraction. Mosbach and co-workers have shown good selectivities for recognizing the "print" molecule over other species with compositional variations as slight as changes in one methyl group. No recognition polymer is going to be 100% selective so redundancy must be built into the sensor array by using templated and untemplated polymer films. HMX, RDX, PETN, DNT and EGDN are distinctively different molecules all having the NO₂ functionality. These species are well-suited for use as "print" molecules via the Mosbach technique. Additionally, the carbonyl group of cyclohexanone can provide the functionality for interactions with MAA when forming a templated polymer to recognize cyclohexanone. Of these compounds, the cyclohexanone is most likely to give the worst results as far as templating because there is only one functional group for hydrogen bonding. Since all of the nitro containing compounds will have high affinities for the polymer films, each SAW sensor may have to be maintained at a different temperature. These temperatures will most likely be above ambient and will be necessary for assuring reversible absorption of the species to be analyzed. The reversibility is required in order to have a dynamic sensor. Thus, the strategy is to have the SAW devices in pairs—one with a templated polymer, the other with an untemplated polymer of the same composition, both at the same temperature. The array would then contain one SAW for humidity, two SAWs each for cyclohexanone, EGDN, DNT, PETN, RDX and HMX. The idea is to provide a dual SAW discriminator for each compound of interest. The templated polymer film should yield a different signal than the untemplated polymer film since all species in the sample stream other than the target species should interact with the two SAWs in the same manner. The dual SAW approach eliminates the need for an "all-or-nothing" response that is not feasible with any recognition surface at this time. The array can be expanded to include recognition of other compounds with minor changes to the second subsystems and no modification of the first subsystem.

D. CONCLUSION

A systems approach to the detection of bombs points out the need for a relatively low-cost distributed front end to a overall detection strategy. Consideration of the explosives, impurities, and additives, and the manner by which trace quantities of explosives are transported lead to a design based on dual vapor-particle detection by templated polymer surface acoustic wave technology.

REFERENCES

1. *Discover*, June 1986, pp. 22-31.
2. *Technology Against Terrorism*, Office of Technology Assessment Report OTA-ISC-481 (1991).
3. *Aviation Week & Space Technology*, Vol. 138 (23), 122, June 7, 1993.
4. Unclassified briefing by Dr. William Dennis, Central Intelligence Agency, August 19, 1993.
5. *Kirk-Othmer Encyclopedia of Chemical Technology*, 3rd ed., vol 9; Wiley-Interscience: New York, p. 583.
6. I.R. Johnston, R.H. Weldon, G.M. Hensel, *Recrystallisation of RDX at W.R.E.*, Tech. Rep. WRE-TR-1760(W), Defence Science and Technology Organisation, Weapons Research Establishment, Salisbury, Australia, February 1977
7. W.S. Wilson, *Recrystallized RDX for RDX/Polyethylene Wax Compositions*, Tech. Note MRL-TN-436, Defence Science and Technology Organisation, Materials Research Laboratories, Melbourne, Australia, May 1980.
8. P.M. Gallagher, M.P. Coffey, V.J. Krukonis, W.W. Hillstrom, *J. Supercrit. Fluids* 5, 130 (1992).
9. G.C. Slack, H.M. McNair, L. Wasserzug, *J. High Res. Chrom.* 15, 102 (1992).
10. A.C. Stern, *Air Pollution*, 3rd ed., vol. 1; Academic Press: New York, 1976, p. 88.
11. Ref. 10, p. 144.
12. Ref. 10, p. 86.
13. M. Corn, F. Stein, *Nature* 211, 60 (1966).
14. "The TWA Experience," in *FAA, Proceedings of the First International Symposium on Explosive Detection Technology*, p. 51 (1992).
15. *Aviation Week & Space Technology*, Vol. 134 (12), 62, March 25, 1991.
16. *Ibid.*, 60, March 25, 1991.
17. "Automatic Detection of Explosives Using X-Ray Imaging" in *FAA, Proceedings of the First International Symposium on Explosive Detection Technology*, 1992, p. 68.
18. *Aviation Week & Space Technology*, Vol. 137 (6), 35, August 10, 1992.
19. *Ibid.*, Vol. 137 (7), 40, August 17, 1992.
20. "Vapor Detection of Explosives," in *FAA, Proceedings of the First International Symposium on Explosive Detection Technology*, 1992, p. 45.
21. *Aviation Week & Space Technology*, Vol. 130 (25), 164, June 19, 1989.

22. Unclassified briefing by Dr. Randy Murch, Federal Bureau of Investigation, August 20, 1993.
23. V.A. Marple, K. Willeke, in *Fine Particles: Aerosol Generation, Measurement, Sampling, and Analysis*, B.Y.H. Liu, ed., Academic Press: New York, 1976, pp. 411-446.
24. V.A. Marple, C.M. Chien, *Environ. Sci. Tech.* **14**, 976 (1980).
25. B.T. Chen, H.C. Yeh, *J. Aerosol Sci.* **18**, 203 (1987).
26. V.J. Novick, J.L. Alvarez, A.D. Appelhans, in *Aerosols*, B.Y.H. Liu, ed., Elsevier: New York, 1984, pp. 143-145.
27. B.W. Loo, J.M. Jaklevic, F.S. Goulding, in *Fine Particles: Aerosol Generation, Measurement, Sampling, and Analysis*, B.Y.H. Liu, ed., Academic Press: New York, 1976, pp. 311-350.
28. W.D. Bowers, R.L. Chuan, T.M. Duong, *Rev. Sci. Instrum.* **62**, 1624 (1991).
29. G. Vlatakis, L.I. Anderson, R. Müller, K. Mosbach, *Nature* **361**, 645 (1993).

**H. TRACING THE ORIGIN OF MYCOTOXIN CBW AGENTS
BY ^{13}C ISOTOPIC FRACTIONATION**

**Peter Chen
Harvard University
Cambridge, Massachusetts**

TRACING THE ORIGIN OF MYCOTOXIN CBW AGENTS BY 13C ISOTOPIC FRACTIONATION

INTRODUCTION

While the value of chemical, biological, and toxin (CBW) weapons against trained and equipped military forces has been questioned, their use against unprotected troops or civilians in either a military or terrorist attack can have devastating results. The Geneva Protocols, and the more recent Chemical and Biological Weapons Conventions ban these weapons, but their production and use can be nevertheless attractive to a State or political organization under the proper circumstances. Key to those circumstances would be the capacity to avoid or frustrate the military or political retaliation ordinarily incurred by the use of prohibited weapons. If a military or political objective could be achieved by the use of CBW agents in such a fashion that sufficient public doubt concerning that use can be maintained, the public political consensus in the United States needed for retaliation can be significantly eroded. It is important to note that it is not at all necessary for the guilty party to conclusively exonerate itself—the ability to plausibly deny a violation of treaty in the face of equivocal evidence to the contrary is enough. Even if more damning evidence is available to the United States government, the confidentiality of sources or methods involved in intelligence gathering may prevent the use of that evidence in the public debate. Particularly well-suited to illegal use are mycotoxin [1] weapons. They are easy to produce in quantity without a sophisticated industrial or scientific infrastructure, they are highly toxic, and they are naturally endemic at low levels in many parts of the world. This paper suggests a purely analytical tool that may be able to distinguish between toxin weapons that are cultured in the laboratory and the otherwise identical naturally occurring toxins.

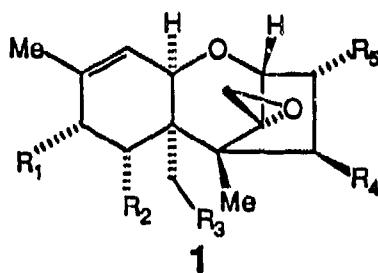
SCOPE

The approach outlined in this paper is general, but the analysis will be done for a representative case, the trichothecene T-2 toxin, that has been illustrative of the difficulties in establishing the prohibited use of mycotoxin weapons. The alleged use of trichothecenes in Southeast Asia and Afghanistan by Soviet or Soviet-client forces was the subject of intense public debate [1] in the late '70s and early '80s, with the ultimate resolution

remaining ambiguous in the public eye. Because trichothecenes that occur naturally in improperly stored grains have caused several large-scale incidents of mycotoxicosis in the past decades [2], even the discovery of T-2 toxin and its metabolites in the affected areas was insufficient to make serious retaliation palatable. This paper will start with a general background to trichothecenes, including a procedure for production of T-2 toxin to convince the reader that simple fermentation methods are easy and the most likely to be used. A section on the biosynthesis of trichothecenes follows as background for the design of passive tracer strategies. Following that will be a discussion on ^{13}C isotopic fractionation in that biosynthesis and the isotopic composition of T-2 toxin that can serve as an indication that the toxin was produced under culture conditions. The final sections will discuss limitations of the method and the general protocol for application to other classes of toxin weapons.

BACKGROUND: TRICHOTHECENE MYCOTOXINS

Trichothecenes are a family of fungal secondary metabolites [3] of the general structure 1, that are produced by common molds of the genera *Trichothecium*, *Cephalosporium*, *Myrohectium*, *Trichoderma*, *Stachybotrys*, *Cylindrocarpon*, *Verticimonosporium*, and *Fusarium*.



The last of these is a common mold that occurs in improperly stored grain. The association of a syndrome called *alimentary toxic aleukia* [2] (ATA) with the consumption of grain infected with *Fusarium* was made subsequent to a massive outbreak of poisoning in the southern Ural region of Russia in 1944. Poisoning is marked by destruction of the skin, hemorrhaging, inflammation, and sepsis, followed by atrophy of the bone marrow and a rapid, fatal drop in leukocyte and erythrocyte count. Protein synthesis is strongly inhibited. Routine monitoring of cereal foodstuffs for mycotoxin contamination is done in the United States and most developed countries. The potency of the T-2 toxin ($\text{R}_1 = \text{isovalerate}$, $\text{R}_2 = \text{H}$, $\text{R}_3 = \text{R}_4 = \text{OAc}$, $\text{R}_5 = \text{OH}$) produced by *Fusarium* molds, as characterized by the LD₅₀ value [4, 5], (50% death upon exposure, in mice), is

3.04 ± 0.14 mg/kg. This makes T-2 toxin about a factor of five more toxic than arsenic and comparable to parathion, a thiophosphate insecticide related to nerve gases.

The total chemical synthesis of several trichothecenes has been reported [6]. While beautiful as examples of the chemist's art, it is unlikely, though, that synthetic trichothecenes will appear as CBW agents unless significantly modified structures show substantially better toxic or delivery properties. Even if they did, though, use of non-natural toxins would surely make it impossible to deny a CBW attack. Natural trichothecenes for use as CBW agents can be easily prepared with rudimentary laboratory equipment and minimal training by growing the appropriate mold in liquid shake culture. The following protocol [5, 7] is reproduced to indicate the ease by which significant quantities of purified T-2 toxin can be prepared. It hardly seems likely that another route would be used because, in contrast to the production of nerve gas, for example, the published fermentation procedures require little chemical infrastructure.

Fusarium tricinctum was grown at 8°C in 500 ml Erlenmeyer flasks containing Gregory's medium (100 ml) for 30 days. The cultures were blended and lyophilized to give a powder (300 g). The powder in 10 g portions was extracted with EtOAc (2×250 ml) and evaporated to afford an oil (20 g), which was dissolved in 100 ml EtOAc and washed with 0.5% H_2SO_4 (3×60 ml). The EtOAc layer was concentrated and 180 ml methanol-water (4:1 by volume) was added. The resulting solution was extracted with Skellysolve B (3×60 ml) and then treated with water (150 ml) to yield a 1:1 methanol-water solution. Extraction of the solution with 1:1 chloroform-EtOAc (3×200 ml) and removal of the solvents in vacuo resulted in a brown oil (6.2 g). Chromatography of this oil on silica gel (465 g) with 6:1 acetone-chloroform, followed by recrystallization from benzene-Skellysolve B yielded T-2 toxin (1.5 g) as white needles, m.p. 151-152°.

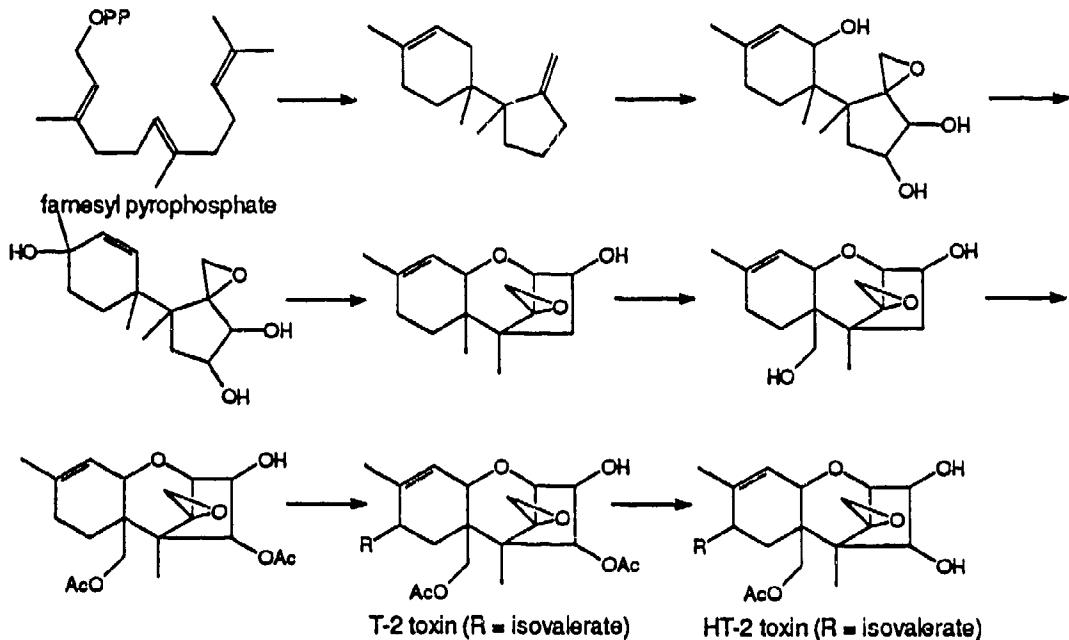
The amazing aspect of the procedure is that the fungus is *at least* 0.5% T-2 toxin by dry weight! Sufficient quantities (a few tens of grams) for a terrorist action could be produced in a high school chemistry lab. Any country with fermentation facilities could produce the multikilogram quantities needed for a military operation. It is also not clear that the extensive purification steps, with associated mass losses, are necessary to make an effective weapon. The oil produced after the initial extraction is probably suitable for use as a CBW agent, especially since the crystalline toxin must be incorporated into some dispersal medium anyways. A notable feature of the preparation is that *growth of the fungus at 24°C instead of 8°C produces a factor of ten less T-2 toxin* [5, 8]. HT-2 toxin, a structure related to T-2 toxin by deacylation ($R_4 = \text{OH}$ instead of OAc), appears in comparable quantities at the higher temperature. The same qualitative result, deacylation at

higher temperatures, has been confirmed for growth of the mold on solid corn, rice, and wheat, with deacylation activity higher for mold grown on solid grains versus liquid culture [3]. The deacylated toxin is several times less toxic [5, 9] than T-2 toxin, and is therefore, not the CBW agent of choice.

The natural occurrence of mycotoxins make them an ideal CBW agent whose use can be publically denied after achieving a desired policy goal. Discovery of a small concentration of a toxin is not *prima faciae* evidence for a violation of treaty. Consumption of bleu cheese exposes the gourmand to a low level of the mycotoxin, roquefortin [2], which is produced by the same *Penicillium roqueforti* mold that gives the cheese its characteristic appearance and flavor. While the LD₅₀ dosage for roquefortin is only 10 mg/kg, the average adult would have to eat 200 kg of cheese to ingest a lethal dose of toxin. A Hmong tribesman in the mountains of Laos is unlikely to eat bleu cheese, but he is likely to eat rice which, even in the United States, can contain low levels of trichothecene mycotoxins if sufficiently sensitive analytical techniques are employed. Therefore, subsequent to a *bona fide* mycotoxin attack, discovery and analysis of actual mycotoxin residues is insufficient proof to make a credible case for political or military retaliation. Additional intelligence can supply the needed proof, but a publically presentable analysis that takes advantage of inherent features in the biosynthesis of a mycotoxin to show that there was a CBW agent would be useful in building the political consensus against the guilty party.

THE BIOSYNTHESIS OF TRICHOThECENE MYCOTOXINS

A handle on the source of T-2 toxin isolated after an alleged CBW incident is provided by the biosynthetic pathway by which the toxin is produced in the mold. Extensive biochemical studies [10, 11] have shown all trichothecenes to derive from the isoprenoid, farnesyl pyrophosphate. The pathway is shown schematically below (without stereochemistry):



With the exception of the two acetate and one isovalerate ester sidechains, all of the carbons in T-2 toxin derive from farnesyl pyrophosphate, which in turn, come from the two-carbon species, acetyl-CoA, by the biosynthetic pathway common to lipids. The two acetates also derive from acetyl-CoA, which leaves only the isovalerate sidechain. Studies on UV-induced mutant strains [11] of the mold have found that the carbons in the isovalerate sidechain come from leucine, which is also traced back to acetyl-CoA. The origin of the carbon atoms in the T-2 toxin is the handle by which one can determine whether the toxin was produced by the large-scale production protocol or not.

TWO STRATEGIES BASED ON ^{13}C ISOTOPIC FRACTIONATION IN T-2 TOXIN

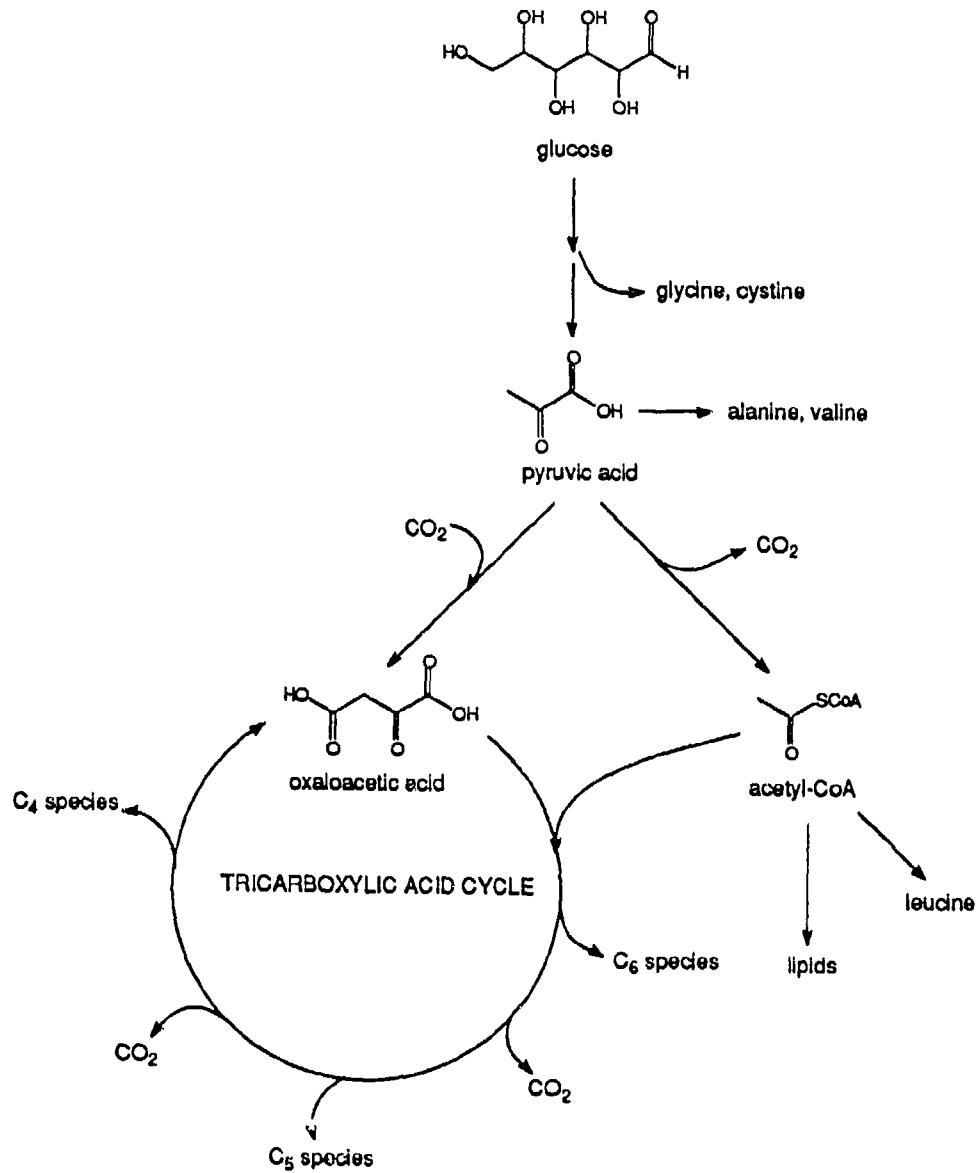
The isotopic abundance of ^{13}C relative to ^{12}C is usually quoted [12] as 0.011, or 1.1%. A more precise measure of the ^{13}C composition of a sample is given by a $\delta^{13}\text{C}$ value which describes the isotopic abundance in parts-per-thousand relative to a standard, usually a fossil limestone called Pee-Dee Belemnite (PDB). With the definition [13]:

$$\delta^{13}\text{C} = \left[\frac{\left(\frac{^{13}\text{C}}{^{12}\text{C}} \right)_{\text{sample}}}{\left(\frac{^{13}\text{C}}{^{12}\text{C}} \right)_{\text{standard}}} - 1 \right] \times 1000$$

a sample with $^{13}\text{C}/^{12}\text{C} = 0.011216$, as compared to 0.011250 for PDB, shows $\delta^{13}\text{C} = -3^{\circ}/\text{oo}$. Isotopic ratio mass spectrometry can determine $\delta^{13}\text{C}$ with an accuracy of $\pm 0.2^{\circ}/\text{oo}$ or better. A negative $\delta^{13}\text{C}$ value means relative depletion; a positive $\delta^{13}\text{C}$ means enrichment.

Rates of reaction depend on the mass of the atoms at or near the bond-making and bond-breaking sites [14]. The effect is small and is often ignored. However, for an intermediate in a multistep reaction that can proceed to two or more different products, i.e., a branching in a reaction pathway, the effect of isotopic substitution on the rates for the two or more pathways will in general not be the same. Therefore, given a statistically random distribution of *naturally occurring isotopes* in the intermediate, there will be a non-statistical distribution in the products after a branching in the reaction pathway. The enrichment or depletion of a minor isotope in one of several possible products by this process is called isotopic fractionation. This section discusses two strategies for identification of cultured T-2 toxin based on ^{13}C isotopic fractionation in the biosynthesis of the toxin by *Fusarium* mold. The first strategy involves the temperature dependence of the overall $^{13}\text{C}/^{12}\text{C}$ ratio in T-2 toxin. The second involves the relative level of ^{13}C enrichment in the acetate group marked R₄ in structure 1. A schematic detailing the incorporation of carbon atoms from glucose into a variety of biosynthetic products is shown below. Because of differing ^{13}C kinetic isotope effects for the two possible reactions shown for pyruvate, one product, acetyl-CoA, is depleted in ^{13}C while the other, oxaloacetate, is enriched [15, 16]. The carbon atoms in lipids derive exclusively from the two-carbon acetyl-CoA unit, so the isotopic depletion is carried through to all lipids and lipid-derived compounds, including the trichothecene mycotoxins. Isotope effects, and hence the depletion or enrichment, is

temperature-dependent. Because cultured T-2 toxin is grown at 8 °C instead of ambient temperature, the $\delta^{13}\text{C}$ value of T-2 toxin is marker of its origin.



In *E. coli*, for example [16], fed with glucose that has $\delta^{13}\text{C} = -9.96 \pm 0.05\text{\textperthousand}$, the overall ^{13}C content of the whole cells is $\delta^{13}\text{C} = -9.67 \pm 0.01\text{\textperthousand}$ while the lipid fraction has $\delta^{13}\text{C} = -16.84 \pm 0.04\text{\textperthousand}$. Because of subsequent chemistry, there is further fractionation within the lipid fraction: the ^{13}C content of fatty acids is $\delta^{13}\text{C} = -12.84 \pm 0.06\text{\textperthousand}$ while that of neutral compounds (including isoprenoids) is $\delta^{13}\text{C} = -25.38 \pm 0.04\text{\textperthousand}$. Feeding studies [15] in which glucose, pyruvate, or acetate was used as the sole carbon source found that the lipids were ^{13}C -depleted relative to

glucose and pyruvate, but unchanged relative to acetate, which shows that the fractionation occurs predominantly at the decarboxylation of pyruvate to form acetate. Importantly, in a study of the ^{13}C content in acetyl-CoA as a function of temperature [15], the $\delta^{13}\text{C}$ value changes by approximately $-20\text{\textperthousand}$ for each rise of $10\text{ }^{\circ}\text{C}$, which should then show up in the products downstream from it. If the same basic biochemistry occurs in *Fusarium*, T-2 toxin cultured at $8\text{ }^{\circ}\text{C}$ should therefore have a measurably different isotopic composition relative to toxin produced on moldy grain at ambient temperature.

The second strategy has the advantage of a larger isotopic fractionation, but it suffers from the disadvantage that the fractionation must be measured at a specific site on the molecule rather than for the molecule as whole. An opportunity for fractionation comes at T-2 toxin itself. In contrast to the T-2 toxin in $8\text{ }^{\circ}\text{C}$ liquid shake cultures, T-2 toxin in moldy grain is only one among many in a mixture of trichothecenes. Overall toxin production is much reduced at room temperature [5], and, importantly, a new trichothecene, HT-2 toxin, is produced by the deacylation of T-2 toxin. Loss of the acetyl group at R₄ of T-2 toxin proceeds slowly at low temperatures but is fast at room temperature. For an ester hydrolysis, a typical ^{13}C kinetic isotope [17] is $k_{12}/k_{13} \approx 1.04$. If the ratio of T-2 to HT-2 toxin is 1:1, i.e., typical for ambient temperature, the acetyl group of the remaining T-2 toxin should be enriched in ^{13}C relative to what it would have been had no deacylation occurred. The enrichment at the carbonyl carbon of the acetyl group at 50% conversion of T-2 to HT-2 toxin is calculated to be $\delta^{13}\text{C} = +28\text{\textperthousand}$ relative to what it would have been at 0% conversion. Because the isotope effect at the methyl carbon is negligible, the enrichment in the acetyl group will be diluted to half that value. Even so, the isotopic enrichment is very large and should be characteristic of naturally produced T-2 toxin. In T-2 toxin cultured at low temperature with negligible deacylation, no relative ^{13}C enrichment at the acetyl group should be observed.

LIMITATIONS ON THE USE OF ^{13}C ISOTOPIC FRACTIONATION AS A TRACER

Any method has limitations and this one is no exception. While the prior work suggests that naturally occurring T-2 toxin can be distinguished from cultured material, there are pitfalls, and also some deliberate countermeasures, which may complicate or frustrate the analysis. The instrumentation for analysis, an isotopic ratio mass spectrometer, is expensive--on the order of \$250,000 for one instrument—but it is commercial off-the-shelf technology. The $\delta^{13}\text{C}$ values cited above were measured in controlled experiments where a single, isotopically homogenous carbon source was given to the microorganism. Field conditions will almost certainly differ from those in the laboratory, but the molds that produce T-2 toxin grow on grains, in which the predominant carbon source is starch, a polymer of glucose that should have the same isotopic composition as glucose. Presumably, in a real-life analysis, the comparison would be made between a sample of T-2 toxin from a putative CBW attack and the same toxin isolated from molds grown on some local grain at room temperature.

A deliberate countermeasure to the first strategy would be the use of blocked mutants of the natural mold to produce the mycotoxin. Beremand and coworkers [11] have used antibody screening to select strains of *Fusarium* mold with UV radiation-induced mutations. Several of the mutant strains accumulate T-2 toxin, presumably because the processing steps in the biosynthetic pathway subsequent to T-2 toxin have been disabled. A judiciously chosen mutant strain may be able to produce large amounts of T-2 toxin at room temperature, which could frustrate the first tracking strategy based on the overall $\delta^{13}\text{C}$ for the whole toxin molecule. The mutants show morphological and other abnormalities which would make them less suitable, in general, for culture than the wild-type mold, but these do not preclude use of the mutants by an adversary intent on countermeasures.

The second strategy, which is based on the $\delta^{13}\text{C}$ value for the acetyl group at R₄ in the toxin, is not affected by the use of a blocked mutant mold. It may, in fact, work a little bit better. The major disadvantage of this approach, though, is the requirement that a site-specific $\delta^{13}\text{C}$ be measured. Usually, whole-molecule $\delta^{13}\text{C}$ values are measured after burning a molecule completely to CO₂. A larger quantity of the toxin would probably be needed for site-specific work, which is a limitation when collection from the field must be done. There have been site-specific $\delta^{13}\text{C}$ measurements in lipids reported [16] in the

literature, so the methodology should work, in principle, for mycotoxins as well. The chemistry preceding the isotopic ratio mass spectrometry must be clean and quantitative. Fortunately, for T-2 toxin, the deacylation that would be needed prior to analysis probably can be done with the yield and specificity needed to make the analysis work. Laboratory trials to show this should be straightforward.

CONCLUSION

The discussion of ^{13}C isotopic fractionation as a tracer for the production of mycotoxin CBW agents has centered specifically on T-2 toxin because of its purported role in "yellow rain" incidents, but also because a fairly complete knowledge of its biosynthesis has emerged in the last two decades. The analysis can be extended to other toxin weapons if the biochemistry is comparably well-known. Especially relevant for another analysis would be the biosynthetic origin of the carbons in the toxin of interest, as well as a knowledge of the pathway immediately after production of the toxin itself. Distinguishing between deliberately produced mycotoxin CBW agents and their natural cousins is difficult, which makes the toxin weapons attractive despite their prohibition, but the demands of large-scale production in culture leave detectable markers that can be used to track their origin.

REFERENCES

1. C.G. Rousseaux, *Comments Toxicol.* **2**, 37 (1988).
2. B. Franck, *Angew. Chem. Int. Ed. Engl.* **23**, 493 (1984).
3. K. Ishii, *Dev. Food Sci.* **4**, 7 (1983).
4. S.G. Yates, H.L. Tookey, J.J. Ellis, H.J. Burkhardt, *Phytochem.* **7**, 139 (1968).
5. Ch. Tamm, M. Tori, *Dev. Food Sci.* **8**, 131 (1984).
6. P.G. McDougal, N.R. Schmuff, *Prog. Chem. Org. Nat. Prod.* **47**, 153 (1985).
7. J.R. Bamburg, N.V. Riggs, F.M. Strong, *Tetrahedron* **24**, 3329 (1968).
8. J.R. Bamburg, F.M. Strong, *Phytochem.* **8**, 2405 (1969).
9. Y. Ueno, *Pure Appl. Chem.* **49**, 1737 (1977).
10. P.S. Steyn, *The Biosynthesis of Mycotoxins. A Study in Secondary Metabolism*, Academic Press: New York, 1980.
11. M.N. Beremand, S.P. McCormick, in *Handbook of Applied Mycology*, vol. 5, D. Bhatnagar, E.B. Lillehoj, D.K. Arora, ed., Marcel Dekker: New York, 1992, pp. 359-384; M.N. Beremand, P.J. Black, *Trichothecene toxins and their production with Fusarium mutants*, U.S. Patent Application No. PAT-APPL-7-173 910, March 28, 1988.
12. H.J.M. Bowen, *Environmental Chemistry of the Elements*, Chap. 10, Academic Press: New York, 1979, pp. 181-193.
13. J.A. Raven, in *The Biochemistry of Plants*, vol. 13, Chap. 4, D.D. Davies, ed., Academic Press: New York, 1987, pp. 127-180.
14. A. Fry, in *Isotope Effects in Chemical Reactions*, C.J. Collins, N.S. Bowman, ed., Van Nostrand Reinhold: New York, 1970.
15. M.J. DeNiro, S. Epstein, *Science* **197**, 261 (1977).
16. K.D. Monson, J.M. Hayes, *Geochim. Cosmochim. Acta* **46**, 139 (1982).
17. M.H. O'Leary, J.F. Marlier, *J. Am. Chem. Soc.* **101**, 3300 (1979); J.F. Marlier, *J. Am. Chem. Soc.* **115**, 5953 (1993).

**I. BASIC SKILLS TRAINING FOR THE CIVILIAN WORKFORCE:
WHAT CAN BE LEARNED FROM THE U.S. MILITARY?**

**Anne B. Myers
University of Rochester
Rochester, New York**

SUMMARY

The problem of workplace literacy is one that has received considerable recent attention. As a result of a combination of generally higher technology in the workplace, rapidly changing job requirements, and slipping educational standards, both young adults (including many high school graduates) and older people currently employed in relatively low-skilled jobs are increasingly seen to lack the basic skills (reading, writing, and mathematics) needed to become or remain productive members of the present and future workforce. The U.S. armed services have also, at least in the past, had to accept fairly large numbers of recruits having deficiencies in these basic skills, and they have developed a variety of remedial training programs in response. This paper asks whether the approaches to basic skills training refined by the military can be usefully transferred to the private sector, particularly in the areas of adult basic literacy and vocational education. Military basic skills training programs emphasize functional (task-based) training methods, which most studies indicate are considerably more effective than general literacy approaches for teaching the skills relevant to a particular job. Military training programs clearly succeed in improving recruits' reading and mathematical skills within a reasonably short period of time, and at least some of those gains are retained for short periods of time, particularly those that are reinforced by regular use on the job. However, evaluations of the longer-term benefits of such training are either not available or equivocal. A largely computerized basic skills training program developed by the Army, the Job Skills Education Program (JSEP), has been revised for civilian applications and is presently being tested at a number of pilot sites around the country. The civilian version of JSEP has been generally well received by students and has been shown to be effective in improving the basic skills of adult intermediate literates over the short term, although data needed to determine whether it is superior to any other approach to basic skills training are not available. The main impediment to the widespread adoption of JSEP or any other computer-oriented basic skills training program seems to be the initial cost of the hardware and software.

I. DEFINING THE PROBLEM

Numerous recent studies document the widely perceived inadequacy of the U.S. workforce in basic skills (reading, writing, and computation) [1-6]. This is viewed as a serious problem affecting both the stability of American society and the competitiveness of the U.S. in the world economy. The increasing presence of technology in the workplace (computers, automation, etc.) and the rapidly changing requirements of many jobs require an increasing level of verbal and computational literacy among workers. Many young people, even those who have graduated from high school, do not possess the skills employers look for in applicants for even the most basic entry-level jobs; it is estimated that one in four high school graduates leaves with the equivalent of less than an eighth grade education [1]. In addition, older employees in relatively low-skilled jobs often find themselves unable to cope with new technologies that enter the workplace. Anecdotes highlighting the inadequacy of basic skills within the U.S. labor pool abound. For example, in 1987 New York Telephone had to process 57,000 applications to find 2,000 qualified entry-level workers [1]. A National Association of Manufacturers survey found that employers usually had to interview six applicants to find one qualified employee, and 30% of companies surveyed said that they could not install new work systems because their employees could not learn new jobs [2]. A 1992 study of job requirements and employee skills at an Arkansas paper mill found that one-third of all employees could not read well enough to perform their jobs safely and half did not have the required level of math skills [6]. Similar problems are shared by other Western industrial nations. One-third of Canadian firms surveyed reported difficulties in introducing new technologies because of the poor basic skills of their workers, officials in Germany estimate that there are 500,000 to 3,000,000 "illiterates" in that nation, and the French Defense Ministry estimated that of men called for military service in 1990-91, 20% could not read adequately [3]. It has been estimated that up to 65% of the entry-level workforce in the United States over the next 15 years may consist of individuals who can read only at the fourth- to eighth-grade level, even though many of these will be high school graduates [1]. Thus it is not surprising that the market value of a high school diploma is falling: the proportion of male high school graduates aged 25-54 whose earnings are below the poverty level for a family of four has grown from 8% to 23% between 1969 and 1989 for whites, and from

20% to 43% for blacks [2]. Some have detected a silver lining in these gloomy statistics. Carnevale suggests that

in some respects, the declining quantity and quality of entry level employees is a happy problem. The scarcity of entry level workers will guarantee work for those who are prepared, inspiring better preparation among people whose prospects have traditionally been limited, and greater willingness among governments and employers to invest in young students and workers [4].

Such a view is even less justified in today's troubled economic times than when it was written several years ago, and in any case, the existence of a large group of functionally near-illiterates in a growingly technological society is troubling at best.

Several recent studies have predicted that the new jobs that will be created during the next 10–20 years will require, on average, a higher level of basic skills than currently existing jobs. In a 1987 report, the Hudson Institute projected the areas of greatest job growth in absolute numbers to be service occupations, managerial and management-related, marketing and sales, administrative support, and technicians, while the fields expected to lose the greatest number of jobs were agricultural, forestering, and fishing, machine setters, operators, and tenders, hand workers, assemblers, and fabricators, miners, and blue collar supervisors [5]. On a scale from 1 to 6, 6 being highest, the fastest growing occupations require an average reading level of 4.2 and math level of 3.1, versus 2.6 and 1.6, respectively, for those jobs predicted to decline in number. Of the new jobs projected to be added between 1984 and 2000, only 4% require only the lowest level of skills, versus 9% of current jobs; 41% require skill levels in the top three categories, compared with 24% in 1984. A 1992 study by the U.S. Department of Labor predicts that the 19 occupations which will provide more than 300,000 new jobs each between 1990 and 2005 include a range of skill levels from high (general managers and top executives, systems analysts and computer scientists, accountants and auditors, computer programmers) to low (janitors, cleaners, and maids, food counter and fountain workers, food preparation workers, gardeners); however, nearly all of the occupations projected to lose workers during the same period fall into the low-skilled category [7]. The Hudson Institute's conclusion that "jobs that are currently in the middle of the skill distribution will be the least-skilled occupations of the future" [5] seems generally reasonable.

The military is the single largest provider of adult education and training in the United States. The armed forces have faced in the past, and continue to face, problems similar to those encountered by business in training young people for jobs that require

higher levels of reading, writing, and/or mathematical skills than they possess prior to entry. Minimal standards for entry into the services rise and fall as force requirements vary, and the quality of new recruits is currently at an all-time high as the drawdown in force size, combined with the economic recession in the civilian sector, allow the armed services to be choosier than ever about who they accept. In fiscal years 1987-1991, 97% of all new recruits into the services were high school graduates [8] (although a high school diploma is by no means a guarantee of literacy as described above), and only 3% scored in the lowest admittable category, Category IV, on the Armed Forces Qualifying Test (described in Section II) [8]. However, the quality of recruits has at many times been far lower, forcing the services to develop programs for remedial training in reading (particularly), mathematics, and other basic skills. In addition, the services train large numbers of people in various technical specialties that have near or exact equivalents among civilian occupations. A casual view from afar suggests that they do a rather good job of such training. Is this indeed the case, and, if so, can general methods or specific training programs developed by the military be successfully and cost-effectively adapted to the civilian workforce?

The organization of this paper is as follows. First, the past and present approaches that the U.S. armed services have taken to identify and train recruits lacking in basic skills are summarized. The emphasis is on reading since it is generally considered the most fundamental vocational skill [1] and has been the major focus of most military remedial programs, although training in mathematics and other skills is also touched on. Second, an attempt is made to evaluate these programs—to determine how well they work, in both the short and long terms, and to address the issue of transferability of skills learned in the military to civilian life. Third, some existing training programs run by companies to improve the basic skills of their employees are reviewed and compared with the military programs. Fourth, one recent ongoing effort to transfer a military basic skills program into the civilian sector is described and evaluated. Finally, the prospects for further development of military training methods and specific programs as useful tools for civilian training are discussed.

II. BASIC SKILLS TRAINING IN THE U.S. MILITARY

The U.S. armed services set minimum physical and mental requirements that must be met by new recruits, whether enlisted or, in prior times, drafted. Since 1950 all recruits have been screened using the Armed Forces Qualifying Test (AFQT), which represents the verbal and mathematical aptitude portions of the more comprehensive Armed Services Vocational Aptitude Battery (ASVAB). Scores on the AFQT are divided into five categories that correlate to percentiles within the national population as follows [9]:

AFQT Category	Percentile
I	93-99
II	65-92
III	31-64
IV	10-30
V	1-9

Each service branch sets its own standards, which also incorporate other factors; for example, non-high school graduates are typically required to have higher AFQT scores than graduates. However, servicewide limits are placed on the proportion of Category IVs that can be taken, and Category Vs are excluded by law [9].

The fraction of Category IV recruits, whose basic literacy skills tend to be weak, has been quite low since the late 1980s, but there have been two periods during which a relatively high percentage of Category IVs were admitted. The first of these was during 1966-1971 under Project 100,000, a program launched by then-Defense Secretary Robert McNamara. Under this program, which was designed both to help meet the manpower needs of the Vietnam War and to provide the supposedly salutary benefits of military training and discipline to disadvantaged youth, approximately 320,000 men who would otherwise not have met existing standards were either drafted or allowed to enlist in the armed services [9]. The second period was during the early days of the present all-volunteer force from 1976 to 1980, when, due to an error in scoring the ASVAB (the "mistraining" incident), more than 300,000 technically ineligible men were unwittingly accepted. (A lively account of how this occurred and why it took so long to be discovered can be found in Ref. 10.) During these two periods, in particular, the military had to

process large numbers of recruits whose academic skills were poor and who were probably, as Laurence puts it, "not very bright individuals" [10].

How do the services turn "low-aptitude" individuals into useful soldiers? Four separate approaches, representing increasing commitments of time and/or cost, can be identified [11]. First, low-aptitude recruits are preferentially assigned to jobs requiring only low skills and little technical training; this appears to be the principal mechanism by which Project 100,000 men were assimilated [11]. Second, extra help or time may be provided in the form of special tutoring or "recycling" recruits through all or part of a training course. Third, training courses and materials may be uniformly revised to make them more "learnable," a process begun as a matter of policy during Project 100,000 and stepped up as a matter of necessity during the ASVAB misnorming incident. Finally, recruits whose basic skills fall below minimal standards may be put through special remedial training courses, usually either before or during basic training; about 12% of the Project 100,000 men received such training [11]. These programs varied in duration according to service from 3 to 8 weeks and were designed to bring recruits to a 5th or 6th grade reading level [10]. While a high percentage of all trainees did reach these nominal reading levels, the longer-term and/or more general benefits of such training are questionable (see Section III).

Each of the four strategies described above can also be identified in the civilian sector. Matching of less skilled job applicants to less demanding jobs occurs as a natural consequence of market forces; unfortunately, in the present recessionary climate the result is more often than the least skilled applicants get no job at all. Extra help and/or time to learn may be provided to some extent in both formal schooling of children and adult education or employer-provided training, although it is clear that at least in the public schools tutors are rarely available and children are rarely failed even if they have not learned the material. A "dumbing down" of standard (non-remedial) courses in primary, secondary, and post-secondary schools is widely perceived and carries over into training programs in the workplace. Finally, explicitly remedial programs in reading and mathematics are becoming increasingly common in businesses, community colleges, and even 4-year colleges and universities.

According to Sticht, the most extensive research on workplace literacy has been carried out by the U.S. military [3]. The services have taken a number of different approaches to literacy training over the years. From World War II to the Project 100,000 era, such training consisted of "general literacy" programs that taught reading in a standard

"academic" way. In the late 1960s, the Army began revising its programs to stress *functional* literacy, using *task-related* rather than "generic" reading materials in teaching reading [3]. This approach is based on research showing that general reading instruction does not necessarily improve performance in on-the-job reading, which tends to emphasize locating information for immediate use rather than remembering content for future reference [12]. The job-specific context also provides "mental hooks" for attaching new information, and it tends to increase motivation to learn since the application is obvious [12]. The Army's Functional Literacy (FLIT) program, initiated in 1971, ran for 6 weeks, 6 hours per day, and was designed to improve a recruit's job-related reading to the same level as someone having a *general* reading level of 7th grade or higher [3]. The Navy instituted two functional-literacy based programs in the early 1980s which incorporate training in other basic skills as well as reading. The Job Oriented Basic Skills (JOBS) program, designed to help lower-aptitude seamen succeed in technical training for which they would not otherwise qualify, includes training in listening, study skills, and mathematics in addition to reading. The Experimental Functional Skills Program (XFSP) is a similar program intended for personnel who have already completed their technical training and wish to improve their chances for promotion, etc. [3]. Both of these Navy programs are still in existence.

The most modern and extensive functional literacy basic skills program provided by the military is the Army's Job Skills Education Program (JSEP) [13]. In 1980, the Training and Doctrine Command (TRADOC) initiated an analysis of 94 of the most common jobs in the Army, defining the basic skills needed to do each job. Beginning in 1983, Florida State University and Ford Aerospace won a contract to develop computer-based lesson plans to help recruits develop these skills. JSEP is the largest computer-oriented basic-skills program ever developed for adults; it contains more than 300 lessons covering over 200 basic skills, takes up over 100 MB of disk space, and incorporates more than 10,000 graphics. About 90% of the lessons are on the computer, the remainder being paper-and-pencil, and all are self-paced. An introductory lesson of about 30 minutes teaches the student how to use the system before starting a preselected lesson plan. Each lesson is graded, and branching to other appropriate lessons is selected depending on performance. Three kinds of skills are taught: reading and writing, computation, and "learning to learn," the latter consisting of time management strategies, motivational skills, reading strategies, test-taking, and problem-solving [13]. A project to transfer JSEP to the domain of civilian adult education, begun in 1988, is the focus of Section V of this paper.

III. EFFECTIVENESS OF MILITARY BASIC SKILLS TRAINING

In view of the large number of individuals who have received basic skills training in the military (the Army in particular), there seems to be surprisingly little hard data addressing the effectiveness of these programs. Most of the extant studies involve woefully small sample sizes and the conclusions tend to reflect the prejudices of the investigator. Here two measures of effectiveness are examined: first, how well do remedial basic skills programs succeed in improving students' basic skills competency, both immediately after completion of training and at later stages in their military careers; second, how well do skills learned in the military transfer into the civilian sector, and, in particular, are low-aptitude veterans better off than individuals of similar ability who never served in the military?

Some studies have indicated that the general literacy programs in place during the World War II and Project 100,000 had fairly impressive success rates, but the evaluation criteria need to be examined carefully. For example, 79% of the Project 100,000 men put in remedial reading courses reached a 5th grade competency within 3-6 weeks, but many of them started out reading very close to that level [10]. Average gains in reading immediately upon completion of the course were $1\frac{1}{2}$ - 2 grade levels, but after 24 months had regressed to only half a grade, perhaps in part because most of these men were assigned to jobs in which little reading was required, so the skills learned were not reinforced [10]. In the early 1970s, the Army's FLIT program was tested at six Army bases and compared with the general literacy course in place at the time. Related functional literacy programs in the Army National Guard and the Air Force, as well as the Air Force's general literacy curriculum, were similarly evaluated. The functional literacy (task-oriented reading) programs averaged three times more gain in job-related reading tasks (2.1 reading grade levels) than did the general reading programs, and, perhaps surprisingly, also benefited *general* reading about as well as did general literacy programs [3]. More recently, low-skilled recruits enrolled in the Navy's JOBS program were evaluated relative to a random comparison group that qualified for technical training w/o remedial basic skills training. The JOBS recruits had considerably greater attrition during technical training

(21% versus 10%), indicating (not surprisingly) that the remedial program did not bring them to a level comparable with the qualified-from-the-outset recruits, but those JOBS recruits who did make it through technical training actually had a lower attrition rate thereafter [3]. Personnel receiving 45 hours of instruction in the Navy's XFSP program outperformed those receiving a comparable amount of general reading instruction in tests of both Navy knowledge and Navy functional skills, but the sample sizes involved were too small to be considered statistically significant [3].

The closest thing to a detailed effectiveness evaluation of the Army's current JSEP program appears to be a study that started with 100 experimental and 100 control personnel matched with regard to sex, race, educational level, and ASVAB and preliminary JSEP test scores [14]. The subjects were soldiers initially testing in AFQT Categories III and IV (see Section II) who had already completed basic training and were enrolled in Advanced Individual Training (AIT) in one of seven occupational specialties for which the Army Research Institute has developed proficiency tests. The experimental group received 36-54 hours of JSEP instruction while the control group received no additional training. Immediately upon completion of AIT, the experimental group scored slightly higher than the controls on the proficiency test; however, after 120 days on the job, only about half the soldiers in each group could be reached for testing, and at that time the experimental group scored slightly *lower* than the controls. The conclusions state that JSEP is effective in both acquiring and retaining job skills [14]. This claim seems to be based on a further analysis which shows that *within the experimental group*, the more hours of JSEP instruction received, the higher the test scores; comparison between experimental and control groups certainly does not support the conclusions.

One of the motivations behind Project 100,000 was the assumption that the training and discipline received in the military would prove beneficial to poorly educated and/or low-aptitude individuals later in civilian life. A number of studies have attempted to determine whether this is true. Sticht, writing in 1987 [11], cited results of a very early (1974) study that found Project 100,000 veterans had a lower level of unemployment and earned more than a matched comparison group of nonveterans. However, a careful study by Laurence *et al.* in 1989 compared a group of 311 Project 100,000 men with 199 controls, aptitude-matched based on test scores from school files, and found that the veterans had consistently lower individual and household incomes [9]. A similar comparison of 326 "potential eligibles" admitted to the services during the ASVAB misnaming with 879 controls, aptitude-matched through ASVAB test scores administered

to the non-veterans, found no significant differences between the two groups in any measure of income [9]. DeTray examined data from 1,702 men aged 14–24 in 1966 who were followed through 1975 and managed to conclude that training for specific military specialties (beyond basic training) increased civilian wages when "innate differences in productivity" were accounted for, even though the data show a consistent *negative* correlation between amount of military training received and civilian wages later [15]. His conclusions were based on a positive correlation between military training and wage *growth* over time after leaving the military [15]. Mangum and Ball presented data on the transferability of military-provided occupational training to the civilian economy, focusing entirely on men and women who served in the post-draft era [16]. They found that income in the civilian workplace was higher among those who received occupational training in the military than for individuals who received all of their job training from business colleges, vocational/technical institutes, correspondence schools, employers, or other civilian sources. Interestingly, when employer-provided training was excluded, the fraction of individuals who ended up employed in the same occupation for which they had been trained was nearly identical (45–50%) for the military-trained and the civilian-trained. This emphasizes how similar many specialties in the modern military are to civilian occupations. Phillips *et al.* recently summarized a large number of studies examining the effect of veteran status on wages as civilians [17]. They concluded that veterans of the World War II and Korean eras generally outearned non-veterans, but Vietnam-era veterans were penalized. The "veteran's premium" for WWII and Korean-era veterans increased with increasing time since serving and tended to be largest for those with the least education and for minorities. On the other hand, the authors' own study, focusing entirely on the all-volunteer force era, found a "veteran's premium" in civilian income only for non-Hispanic whites, and not for minorities [17].

In conclusion, then, the effectiveness of military basic-skills education programs remains difficult to evaluate. It seems clear that functional-literacy programs can significantly improve narrowly defined, task-specific reading and computational abilities in a fairly short period of time, and may help general literacy as much as general reading programs do, but longer-term benefits of such training are not documented. The many studies examining the effect of veteran status on economic success as civilians are themselves equivocal and are of little help in answering the question posed here since they do not distinguish veterans who received special basic-skills training from those who did not.

IV. BASIC SKILLS TRAINING IN THE WORKPLACE

Basic literacy training, broadly defined, has many different providers within the civilian sector. These include public and private primary and secondary schools, community colleges, various other non-profit and for-profit adult education courses, government programs, and employers. No attempt will be made here to comprehensively review either the methods used by these various providers or their effectiveness; instead, discussion will be limited to a cross-section of employer-provided basic literacy programs.

A 1991 evaluation of 29 workplace literacy projects concluded that the most effective programs are those which incorporate systematic analysis of job literacy requirements and development of instructional materials related to these requirements [18]—i.e., a functional literacy approach similar to that characteristic of current military programs. A 1990 report surveyed 19 companies that provide some level of basic skills training to their employees [19]. Many of these programs were initiated when new technologies brought into the workplace revealed the inadequate basic skills of employees whose performance had previously been acceptable. The magnitude and style of these programs vary widely, but a few generalizations can be stated. Large companies tend to use in-house staff for training, while smaller ones contract with community colleges, local school systems, or nonprofit literacy groups. Almost all courses are taught on-site as a way of emphasizing the employer's commitment to training and to minimize inconvenience to employees. Classes are usually conducted on the employees' own time or split between company and employee time, with 3–4 hours per week being common, although some companies simply provide individual tutoring or self-paced courses on a drop-in basis. Both fixed length and open-ended courses are offered. Incentives for participation include increased opportunities for promotion and, in some cases, an understanding that an employee's present job may be in danger if he does not upgrade his skills. Recognition for achievement is a delicate issue since many employees are embarrassed to admit they need remedial training in the first place. In this regard, basic-skills courses are often given names that somewhat disguise the nature of the course, such as "Effective Business Skills" (Aetna) or "Technology Readiness" (Polaroid) [19].

While most of the company representatives surveyed were fairly enthusiastic about the prospects for computer-aided training, only 1 of the 19 companies was using computers for regular instruction, with several others making some use of computers for reinforcement and drill. The primary reasons for the relatively low use of computer-aided instruction seem to be inertia, cost, difficulty in scheduling time on terminals already in use, and a feeling among some that a human interface is required for effective learning [19]. Unless the situation in 1993 has changed drastically from that in 1989, it seems that most employer-provided literacy programs rely largely on non-computer-based instruction. So, in fact, does much military training; although the Army's JSEP program is almost entirely computerized, as recently as 1989 only 2.9% of all individuals receiving formal training in the Navy were in self-paced (computerized or otherwise) courses [20]. It is therefore appropriate to summarize a few relevant points from the mountain of literature addressing the relative merits of computer-based versus "traditional" instruction. A consistent conclusion of many studies is that when instructional time is held constant, students in conventional classrooms with 30 or so students to one instructor score approximately two standard deviations lower than those given individual one-on-one tutoring [21]. Another widespread finding is that when the goal is to bring students to a certain level of proficiency in the minimum amount of time, students using self-paced interactive courseware typically require 20–30% less time for instruction, on average, than those taught in a traditional group "lock-step" manner [20]. This implies that over the long term, interactive computer-based instruction should become cost-effective in settings where training time costs money, which include both the military and businesses. Thus far, however, the cost of the initial investment in hardware and software has apparently acted to override the eventual gains that computer-based instruction might achieve down the road. A final benefit of computer-based instruction for basic skills training is that it can help eliminate the stigma associated with such training (students can tell their peers that they are attending "computer class" not remedial reading) [13], and can incidentally teach some basic *computer* literacy to workers who may have never used a computer before.

V. TRANSFER OF A MILITARY TRAINING PROGRAM TO THE CIVILIAN SECTOR: JSEP

The possibility that military basic skills education has something to offer to the civilian sector is not a new idea. The U.S. Department of Education funded Florida State University and Ford Aerospace to modify the Army's JSEP program (described in Section III) to make it appropriate for use in civilian occupational training, and it is presently being tested on a small but significant scale. As of early 1992, 11 demonstration sites had been established in industrial settings, community colleges, and several other locations [22]. The first two were at Meridian Community College/Peavey Electronics Corporation in Mississippi and at the White Plains Continuing Education Center in New York state. The participants in the preliminary Mississippi demonstration were 64 industrial workers employed by Peavey Electronics who tested at a 5th to 8th grade reading level [23]. The first group of White Plains participants, of unspecified number, were mostly unemployed adults typically reading at a 4th to 6th grade level [19]. An evaluation of the White Plains project conducted by the Department of Education found average improvements of 1.26 grade levels in reading and 0.94 grade levels in math after an average of 84 hours of JSEP instruction [24, 25]. Unfortunately, JSEP has evidently not been tested against any *alternative* basic skills program, nor has long-term retention been evaluated. At the Meridian site, supervisors were surveyed to determine what effects they thought JSEP training was having on workers. They typically reported some improvement in both quantity and quality of work, job attitude, and job knowledge, but these are very qualitative conclusions [23].

A National Alliance of Business report discussed the overall impressions of the JSEP program from students and instructors at both the Meridian and the White Plains test sites [19]. The students almost uniformly liked JSEP, found it interesting, felt the material they learned was useful both on the job and in their everyday lives, and found learning to use the computer quite easy. The instructors' comments were somewhat less rosy, probably due in large part to the number of technical problems encountered in both the hardware and the software as delivered; the test sites had been expecting a product at a commercially marketable level of sophistication, which is not what they received.

Elimination of glitches and more readily available technical support (e.g., a toll-free telephone number) were cited as major needs for the product to become commercially viable. Instructors also felt that the computer-based lessons should be supplemented with some classroom work designed to teach listening, speaking, and vocabulary skills. The continued presence of much Army-based vocabulary and content was cited as a drawback to the general use of JSEP in the civilian community (one lesson, for example, described how to kill, clean, and cook rodents). These aspects are being corrected in updated versions.

Cost was cited as a major hurdle to the adoption of JSEP by civilian organizations. The version tested in the pilot studies utilized much proprietary hardware and software and cost \$40,000 for the host computer and \$6,100 for each student workstation [19]. However, as of 1990 the program was being revised to allow standard inexpensive personal computers to serve as workstations, attached through a local area network to a host computer. A potentially more serious problem seems to be that in order to retain the functional context on which the JSEP program is based, it should ideally be customized to some extent for each user, or at least for each type of user [19]. Customization would certainly increase the cost of the product and could lead to considerable variation in the JSEP's delivered to different end users, with concomitant problems in technical support. Copyright and legal issues, which were cited as ongoing problems at the time of the pilot projects [19], have since been largely resolved [25].

VI. CONCLUSIONS AND PROSPECTS

It seems clear that appropriate training can lead to significant improvements, within a fairly short period of time, in the job-relevant reading and mathematical skills of many relatively low-skilled but reasonably motivated adults, and that with continued reinforcement those gains are retained to at least some extent. Approaches that stress task-related functional context are found to be more effective in developing job-relevant skills than are more general, academically oriented courses, and self-paced instruction, most conveniently delivered by computer, is most efficient in terms of the time required to reach a certain level of proficiency. The armed forces have been using functional literacy approaches in their basic skills training for some time, and while computer-based instruction is still rather sparsely used in the military as a whole, the Army has developed and tested a basic skills program (JSEP) that is almost entirely computer based. JSEP has been modified for civilian use and has been found to be qualitatively effective in adult education and employer-provided basic skills training, but evidently has not been tested in any controlled way against alternative methods or other computer-based packages. The usefulness of any such product will ultimately depend on its cost-effectiveness. The early implementations of JSEP were extremely expensive due to the platform on which it was delivered, but there is no obvious reason why it cannot be adapted for use on fairly inexpensive personal computers, and the potential market for such a product seems large enough to make its development commercially viable. Some compromise will be required between the ideal level of customization, which would involve developing a separate version for each user, and the economically practical level.

Although this paper has focused on the basic skills of reading and mathematics, the armed services have also developed a number of other computer-based training programs that have potential portability to civilian job training. Fletcher identified, as of late 1992, a total of 2,718 interactive courseware programs having possible utility in private-sector training [26], spanning an enormous range of subjects from basic academic and workplace skills to highly specialized technical training relevant outside the military. Some of these programs, if transferred to the private sector, may also prove useful in meeting non-military workforce training requirements.

Throughout this paper, the emphasis has been on corollaries between military basic skills training and civilian adult vocational training. These have much in common: in both situations the basic skills are viewed as a prerequisite to performing a well-defined task, and the goals of the training can be narrower (and correspondingly easier to achieve) than in primary and secondary-school education where general literacy is the goal. Both military training and employer-provided job training emphasize achieving a certain minimal level of proficiency in the shortest possible time, while primary and secondary school education usually operate within a predefined time period. Finally, motivation levels must be a major factor. Much of what ails U.S. primary and secondary schools would probably vanish if all of the students showed up actually wanting to learn. Recipients of military and civilian adult education, however disadvantaged their backgrounds, tend to have fairly strong economic incentives to learn, which the use of obviously job-relevant training materials reinforces. All of these comparisons suggest that military-style training methods should find far more application in adult vocational education than in grade schools or high schools. On the other hand, the use of computers as a teaching tool may alleviate some of the boredom that turns many students off, and if the affinity of many children and young adults for video games is any indication, boredom may not set in once the novelty wears off.

While the "three R's" are certainly important both to America's economic competitiveness and to the earnings potential of individuals, they constitute only a small subset of the skills that are needed in the modern workplace. "Thinking" skills (the ability to learn, reason, think creatively, make decisions, and solve problems) and personal qualities (responsibility, self-esteem, self-management, sociability, and integrity) are at least equally important [2]. Indeed, one could argue that the availability of such aids as hand-held calculators and spell-checkers makes the basic skills *relatively* less important than they used to be. With the rapid pace of change in the workplace and the world in general, employees cannot expect to spend all their working lives doing essentially the same job, or even working for the same employer. Flexibility and the ability to learn are of utmost importance, and there is little evidence that the military, or anyone else, has developed particularly good approaches to teaching these skills. They are perhaps more the province of "education" (development of a broad background of knowledge and critical thinking abilities) rather than "training" (teaching performance of specific tasks). So, while military-derived basic skills training methods may prove useful in remediating some of the flaws in our present educational and social system, they are clearly not enough.

REFERENCES

1. Carnevale, Anthony P., Leila J. Gainer, and Ann S. Meltzer, *Best Practices: What Works in Training and Development (Basic Skills)*, Alexandria, VA: American Society for Training and Development, 1989.
2. The Secretary's Commission on Achieving Necessary Skills, U.S. Department of Labor, *Learning A Living: A Blueprint for High Performance*, Washington, DC: U.S. Government Printing Office, April 1992.
3. Sticht, Thomas G., *The Military Experience and Workplace Literacy: A Review and Synthesis for Policy and Practice*, El Cajon, CA: Applied Behavioral and Cognitive Sciences, Inc., October 1992.
4. Carnevale, Anthony, *America and the New Economy*, Alexandria, VA: American Society for Training and Development, 1990.
5. Hudson Institute, *Workforce 2000: Work and Workers for the 21st Century*, Washington, DC: U.S. Government Printing Office, June 1987.
6. Martin, Joann C., *A Study of the Required, Self-Perceived and Assessed Basic Skill Needs for Personnel Within a Paper Mill Industry*, Ed.D. Dissertation, University of Arkansas, 1992.
7. U.S. Department of Labor, Bureau of Labor Statistics, *Outlook 1990-2005*, BLS Bulletin 2402, Washington, DC: U.S. Government Printing Office, May 1992.
8. Sterzinger, James J., *Education: A Defense Industry in Transition*, ICAF-FAP Report NDU-ICAF-92-S85, Washington, DC: Fort Lesley J. McNair, April 1992.
9. Laurence, Janice H., Peter F. Ramsberger, and Monica A. Gribben, *Effects of Military Experience on the Post-Service Lives of Low-Aptitude Recruits: Project 100,000 and the ASVAB Misnормing*, HumRRO Final Report FR-PRD-89-29, Alexandria, VA: Human Resources Research Organization, December 1989.
10. Laurence, Janice H., and Peter F. Ramsberger, *Low-Aptitude Men in the Military: Who Profits, Who Pays?* New York: Praeger, 1991.
11. Sticht, Thomas G., William B. Armstrong, Daniel T. Hickey, and John S. Caylor, *Cast-Off Youth: Policy and Training Methods from the Military Experience*, New York: Praeger, 1987.
12. Philippi, Jorie W. "Matching Literacy to Job Training: Some Applications from Military Programs," *Journal of Reading*, 31 (7) (April 1988), pp. 658-666.
13. Wilson, Lois S., "An On-Line Prescription for Basic Skills," *Training and Development Journal*, 44 (4), (April 1990), pp. 36-41.
14. Kinnison, William E., *Job Skills Education Program (JSEP) Test Plan Study: Effects of JSEP on Acquisition and Retention of Job Skills*, TRADOC final report, Ft. Monroe, VA: TRADOC, August 1990.
15. De Tray, Dennis N., *Veteran Status and Civilian Earnings*, Rand Corp. interim report R-1929-ARP, Santa Monica, CA: The Rand Corporation, March 1980.

16. Mangum, Stephen L., and David E. Ball, "The Transferability of Military-Provided Occupational Training in the Post-Draft Era," *Industrial and Labor Relations Review*, 42 (2) (January 1989), pp. 230-244.
17. Phillips, Robert L., Paul J. Andrisani, Thomas N. Daymont, and Curtis L. Gilroy, "The Economic Returns to Military Service: Race-Ethnic Differences," *Social Science Quarterly*, 73 (2) (June 1992), pp. 340-359.
18. Kutner, Mark A., *A Review of the National Workplace Literacy Program*, Washington, DC: Pelavin Associates, Inc., May 1991.
19. Frederick, F., C. Fisher, G. Moore, J. Philippi, and V. Rebata, *Lessons Learned: Job Skills Education Program Final Report*, Washington, DC: National Alliance of Business, Inc., May 1990.
20. Angier, Bruce N. and J.D. Fletcher, *Interactive Courseware (ICW) and the Cost of Individual Training*, IDA Paper P-2567, Alexandria, VA: Institute for Defense Analyses, November 1992.
21. Fletcher, J.D. *Individualized Systems of Instruction*, IDA Document D-1190, Alexandria, VA: Institute for Defense Analyses, July 1992.
22. Branson, R.K., "Technology Transfer and the Job Skills Education Program: Preliminary Results," paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
23. Philippi, Jorie W., *Technology Transfer Partnership Project: Meridian Community College-Peavey Electronics Corporation Lessons Learned Report*, report submitted to the National Alliance of Business, January 1989.
24. U.S. Department of Education evaluation report, cited by Lois S. Wilson, personal communication.
25. Paul Geib, U.S. Department of Education, personal communication.
26. Fletcher, J.D., Ruth Wienclaw, Gary Boycan, James Bosco, and Harold F. O'Neil, Jr., *Defense Workforce Training Programs*, IDA Paper P-2743, Alexandria, VA: Institute for Defense Analyses, October 1992.

**J. BENEFITS OF INCREASED AUTOMATION FOR
NUCLEAR NAVAL REACTOR OPERATION AND
PERSONNEL TRAINING**

**Gerald A. Navratil
Columbia University
New York, New York**

EXECUTIVE SUMMARY

Given the relatively limited performance of digital computer technology in the 1950s and 1960s as well as the need to rapidly field nuclear submarines to meet a growing Soviet threat, the Nuclear Navy under the leadership of Adm. Rickover minimized the level of automated systems in the navy's nuclear propulsion systems as a matter of policy. This policy derived from an engineering philosophy which had as its central premise that only established technology should be employed and that highly developmental or untried concepts should not. This philosophy was extended to include an absolute ban on the use of computer-based simulators for the training of nuclear naval reactor personnel. Until very recently this has been the governing philosophy of the nuclear navy and virtually all training of nuclear naval reactor personnel was carried out in shore-based operating reactor prototypes, and the engineering spaces on U.S. nuclear submarines have the lowest level of automation of all the systems on board.

Today both the strategic situation and the level of computer technology are radically different from what existed in the 1960s when the nuclear navy's engineering philosophy was established by Adm. Rickover. With the collapse of the Soviet Union, the need to rapidly field large numbers of ever more capable submarines to meet a growing Soviet threat has been removed. With the tremendous advances in microprocessor capability coupled with the large experience base using high fidelity simulators for training civilian nuclear reactor operators which was mandated by Congress in 1982, the use of automated systems for control and for simulator training can no longer be regarded as a "highly developmental or untried concept." The profound change in the strategic situation has led the Navy to be driven by a need to reduce costs of operations while maintaining an effective submarine fleet. The design and construction of new, more effective submarines is no longer driven by the need to meet a rising threat, but rather to preserve the industrial base which supports the nuclear navy. In light of these greatly changed circumstances, a reexamination of the original engineering philosophy of the nuclear navy is certainly warranted. In this report we discuss the rationale for the Nuclear Navy to change its policy regarding the use of automation for power plant operation aboard submarines and the use of simulators in the training programs of nuclear naval reactor personnel. To implement

this change in policy, we make the following specific recommendations for action by the Nuclear Navy:

- An integrated training program involving both simulator and operating reactor systems should be established for initial and recurrent training of officer and enlisted personnel.
- The development of a nuclear reactor simulator station for installation on board submarines and nuclear surface ships should be pursued. Since studies have shown a significant reduction in the timely crew response to emergency situations after months of largely routine operation, such a ship-based simulator would be used as a training device to ensure a high level of crew readiness for emergency transient response. Additionally, the computing power available in state-of-the-art simulators is now sufficient to provide an analysis tool for the guidance of engineering officers during the evolution of an actual serious reactor casualty. Such ship-based recurrency training may also greatly reduce the need for shore-based training for experienced crew members.
- The introduction of high fidelity simulators of the reactor and engineering control panels in nuclear navy training centers should be exploited as a laboratory to:
 - Develop improved panel layout for minimization of operator errors and improved crew coordination based on quantitatively documented data obtained during actual crew training
 - Test various levels of automation of the engineering spaces in submarines and assess implications on crew requirements
 - Explore the possibility of connecting the reactor control simulator stations to the existing command and fire control simulators to allow full scale exercises integrating engineering and tactical response under normal and various levels of casualty conditions.
- Given the changed strategic situation which substitutes preservation of the industrial base (rather than meeting an immediate threat) as the primary motivation for new submarine production, the Navy should consider this an opportunity to build several prototypes before committing to large scale production of a replacement for the Los Angeles class and older SSNs in the fleet. One objective of this series of prototypes would be to develop the technology needed to bring the automation level in the engineering systems up to the current state of the art, leading to a smaller and less costly nuclear attack submarine.

BENEFITS OF INCREASED AUTOMATION FOR NUCLEAR NAVAL REACTOR OPERATION AND PERSONNEL TRAINING

INTRODUCTION

When the U.S. Navy committed to developing nuclear powered ships shortly after the Second World War, it was faced not only with the need to develop the technology required to design and build nuclear propulsion systems, but also with the need to invent a new engineering culture whose level of technical perfection greatly exceeded existing standards in the Navy due to the extreme hazard inherent in the use of nuclear fission. The story of how this engineering culture was created under the leadership of Adm. Rickover has been described in great detail in a number of recent books [1, 2, 3]. In this paper we examine one aspect of that engineering design philosophy: a strong aversion to automation of any kind and an absolute prohibition of automated simulators in training. This element of Rickover's nuclear submarine design philosophy was arguably the optimal one given the strategic and technological situation he faced. However, the profound changes in the strategic situation brought on by the collapse of the Soviet Union and the tremendous technological advances in microprocessor and computer memory capability which have occurred in the past 20 years certainly call for a reexamination of this element of his philosophy.

We will argue in this paper that it is in the navy's best interest for a substantial modification of this element of the nuclear navy's traditional engineering design philosophy from one of deep skepticism to enthusiastic embrace. The implications of such a change would be significant for the future of the nuclear navy since this aspect affects the design and size our future nuclear-powered ships as well as how the men and women who will run those next generation of ships will be trained. In this paper the focus will be on the introduction of simulators into the training program for nuclear navy personnel. Once it is accepted that a computer-based system is capable of accurately reproducing the normal and abnormal behavior of naval nuclear propulsion systems in real time, the technical foundation has been laid for the addition of further levels of automation into the design of future propulsion systems.

The organization of this paper consists first of a brief review of the traditional engineering philosophy of the U.S. nuclear navy and the training facilities used for nuclear

navy personnel. After a discussion of the capabilities of state-of-the-art nuclear power plant simulators together with a review of the experience base from both civilian (U.S. nuclear electric power utilities) and military (UK Royal Navy and French Navy) use of simulators in training nuclear reactor operators, the issue of engineering space automation on submarines and the Centurion design is discussed. The paper concludes with a discussion and set of recommendations on a course of action for the Navy.

Nuclear Navy Engineering Philosophy

The best synopsis of the engineering philosophy comes directly from the originator himself. The following is excerpted from Adm. Rickover's testimony to Congress in 1982 where he outlined the engineering philosophy of the nuclear navy including 11 important tenets:

Because a warship must be able to perform its mission and return under combat conditions, the nuclear propulsion plant therefore must be engineered to survive battle damage and severe shock; to operate reliably and safely in close proximity to the crew; and to be repaired at sea by the crew if necessary. Standards for materials and systems are rigorous and only premium products with a proven pedigree are used in the reactor to minimize maintenance and take maximum advantage of long core lives.

Building and operating effective naval nuclear propulsion plants involves many engineering and design considerations. The following are important tenets of the program's engineering philosophy:

- Avoid committing ships and crews to highly developmental and untried systems and concepts.
- Ensure adequate redundancy in design so that the plant can accommodate, without damage to ship or crew, equipment or system failures that inevitably will occur.
- Minimize the need for operator action to accommodate expected transients. If the plant is inherently stable, the operator is better able to respond to unusual transients.
- Simplify system design so as to be able to rely primarily on direct operator control rather than automatic control.
- Select only materials proven by experience for the type of application intended and insofar as practicable, those that provide the best margin for error in procurement, fabrication, and maintenance.
- Require suppliers to conduct extensive accelerated life testing of critical reactor systems components to ensure design adequacy prior to operational use.

- Test new reactor designs by use of a land-based prototype of the same design as the shipboard plant. Prototype plants can be subjected to the potential transients a shipboard plant will experience, so problems can be identified and resolved prior to operation of the shipboard plant.
- **Train operators on actual operating reactors at the prototypes. Simulators are not an acceptable training device for naval operators.**
- Confirm reactor and equipment design through extensive analyses, full-scale mockups, and tests.
- Use specially trained inspectors and extensive inspections during manufacture; accept only equipment that meets specification requirements.
- Concentrate on designing, building and operating the plants so as to prevent accidents, not just cope with accidents that could occur. [4]

The fourth and the eighth tenet (highlighted in bold face above) offered by Adm. Rickover derives from his long held belief that automation in all forms was highly undesirable. In his earlier testimony to Congress in 1965 which was then investigating the loss of the SSN *Thresher*, he stated that current American nuclear submarines, in his opinion, are overloaded with automation devices, which unwarrantedly increases their construction cost, lowers their reliability, and makes their operation more complicated [5].

If we examine the technical and strategic context in the period from 1955 to 1970 when Rickover's engineering philosophy was developed and implemented, the key factors governing the evolution of his philosophy were:

- Nuclear energy was a new technology demanding a vastly higher technical standard in construction and operation than existed in the U.S. Navy at that time.
- The capability of automated systems was limited and low MTBF (mean time between failures) for components led to relatively poor reliability of complex systems.
- The rapid deployment of SSNs and SSBNs was needed to match the capability of the expanding Soviet submarine fleet and to quickly establish the third leg of the nuclear defense triad which promised full survivability to a nuclear first strike.

In view of these factors the engineering and design strategy which strongly discouraged the use of automation in training and nuclear power system design was not only highly effective, it was probably optimal for this period of history.

Nuclear Navy Current Status and Training Requirements

As background information for the later discussion of the future needs in training and capability in the U.S. Nuclear Navy, Table 1 summarizes the U.S. Navy nuclear fleet for 1992 as reported in the 1993-94 issue of Janes.

Table 1. 1992 U.S. Nuclear Ships

Type	Class	Active	First Active	Power (khp)	Displace-ment (tons)	Speed (kn)	Crew
SSBN	Ohio	13+ (5)	1981	60 (S8G)	18,770	20	155
SSBN	Ben Franklin	10	1964	15 (S5W)	8,250	25	143
SSN	Centurion	proposed	-2000	?	?	?	?
SSN	Seawolf	(2)	-1996	52 (S6G)	9,137	35	133
SSN	Los Angeles	52+(10)	1976	35 (S5W)	6,927	32	133
SSN	Sturgeon	32	1967	15 (S5W)	4,960	30	120
SSN	Narwhal	1	1969	17 (S5G)	5,830	25	120
Total Active: 108 with 17 under construction ()							
Additionally, other nuclear surface ships consist of							
6	Nimitz Class	CVNs	2 PWR	A4W/A1G	260 khp		
1	Enterprise	CVN	8 PWR	A2W	280 khp		
4	Virginia	CGN	2 PWR	D2G	70 khp		
1	Truxtun	CGN	2 PWR	D2G	70 khp		
1	Bainbridge	CGN	2 PWR	D2G	70 khp		
1	Long Beach	CGN	2 PWR	C1W	80 khp		

If we use the reported crew size accounting for the fact that all SSBNs are supported by two complete crews which alternate shore based and sea patrol duty, the 1992 fleet was supported by about 17,800 active crew members. Since typically about 40% of the crew in a U.S. nuclear submarine is part of the engineering section, we can estimate that nuclear training is needed to sustain a complement of about 7,000 people.

Historically, all of these people received their primary training through the use of land-based prototype reactors. In most cases these were built as full-scale mock-ups of the nuclear submarine engineering space complete with turbines, auxiliary equipment, and shaft power driving a propeller in water as a load. There are currently nine operating land-based prototypes at three locations as summarized in Table 2.

Table 2. Land Prototypes

Reactor Type	Start	Initial Installation Class	Location
S1W	1953	Nautilus	Idaho
S1G	1955	Seawolf	West Milton
A1W (2)	1958-59	Enterprise	Idaho
S3G	1958	Triton	West Milton
S1C	1959	Tullibee	Windsor
D1G	1962	Bainbridge	West Milton
S5G	1965	Narwhal	Idaho
MARF	1976		West Milton
S8G	1978	Ohio	West Milton

These prototypes were used for engineering development, testing of new components and instrumentation, and training. In the early years of the naval reactor program, these prototypes were an integral and essential part of the development program of new reactor power plants for the fleet. However, as the state of the art has matured, it was no longer necessary to actually build a prototype before committing to the construction and deployment of a new reactor as shown clearly by the absence of prototypes for the S5W, S6W and S6G class of reactors which power the large majority of the current nuclear submarine fleet. These facilities continue to play an important role in the initial training of both officers and enlisted personnel assigned to the engineering section of nuclear vessels.

NUCLEAR REACTOR CONTROL ROOM SIMULATORS

While the nuclear navy has followed a philosophy which strongly resisted the introduction of automated systems both in power plant design and in training, other parts of the nuclear power field have followed a very different path regarding the use of automated systems. In this section we briefly summarize the use of nuclear power plant simulators in the U.S. civilian nuclear power industry and in the navies of the UK and France.

Civilian Nuclear Power Experience

The development and implementation of nuclear power plant control room simulators for training in the civilian nuclear power industry parallels the developing capability of commercially available computing hardware. The first training simulator was built for a boiling water reactor (BWR) in 1968 by General Electric followed by the first simulator built for a pressurized water reactor (PWR) in 1970 by Singer-Link. The first simulator purchased by an electric utility was in 1973 by Consolidated Edison for its PWR

plant at Indian Point in New York. In the 1970s most simulator training of electric utility control room personnel was provided by the two primary reactor suppliers, GE and Westinghouse, at central facilities. It was not viewed as generally cost-effective for individual utilities to purchase their own simulators.

The situation regarding the use of simulators radically changed for the civilian nuclear power industry after the 1979 accident at Three Mile Island (TMI) focused attention on the critical importance for safety of operator procedures and training. The TMI accident led to formulation of a TMI Action Plan by the U.S. Nuclear Regulatory Commission (NRC) to develop a methodology and regulations to improve training of reactor operators. This effort by the NRC was preempted by the U.S. Congress in the 1982 Nuclear Waste Policy Act which mandated the NRC to "establish simulator training requirements" for the civilian nuclear industry. Since the NRC had not had enough time to develop its own approach to qualification and use of simulators in training, the NRC issued rules based on the Federal Aviation Administration (FAA) model which emphasizes fidelity standards for simulator equipment used to qualify personnel. These rules led to the widespread purchase by electric utilities in the U.S. of "plant-specific," high-fidelity simulators for training operators and for operator license examinations as codified by the 1985 ANSI/ANS Standard 3.5.

This regulation driven demand for high-fidelity simulators coincided with the phenomenal growth in performance and reduction in cost of computing capability through the 1980s and continuing undiminished through to the present. Consequently the capability and fidelity of the current generation of simulators provided by simulator manufacturers like Westinghouse, S3 Technologies, and CAE, are very substantially higher than anything available in the 1970s.

At the present time, the current state-of-the art simulators use two energy groups to model the reactor core and two phase thermal hydraulic models for the primary and secondary coolant systems in a PWR. These simulators provide excellent fidelity for all normal operations. The simulator facility itself is an exact, full scale copy of the specific plant control room and nearby supporting systems. The instructor can independently control (i.e., fail) anything in the plant: every valve, every indicator bulb, every gauge. Full playback for debriefing in both accelerated, slowed, and real time rates is available. This includes multiple valve faults: open, closed, seeping, fluttering, and partial obstruction.

The current generation of simulators provide adequate fidelity for hypothetical FSAR (Final Safety Analysis Report) accident situations which have been analyzed in detail by mainframe-based industry standard computer codes. However, the capability in current simulators to predict new, unanalyzed events is very limited and given the absence of an validated data base for comparison. The much high power of the next generation of processors (DEC Alpha, Intel Pentium, etc.) should allow analysis capability in control room simulators, but few electric utilities are likely to spend the money needed to validate this capability.

The cost of acquisition of a new plant-specific simulator is currently about \$10–15 million, with operations cost typically \$0.5 million for software upgrades and a maintenance staff of about 5 FTE people. Relative to the multi-billion dollar replacement cost of the plant and several hundred million dollars of electricity sales per year, this cost is modest.

The basic training of civilian reactor operators consists of roughly 8 weeks on the simulator plus 13 weeks in the nuclear power plant. Requalification for operators requires from 40 to 80 hours per year on the simulator. The emphasis of the simulator training is on accident and off-normal scenarios and team response in the control room. It would be uneconomic and in most cases highly unsafe and illegal to attempt to demonstrate this class of scenarios for emergency response training using an operating nuclear reactor.

After more than 10 years of experience with the widespread use of simulators for civilian nuclear power operator training, some general conclusion can be drawn. It is clear that only simulators can provide realistic training experience for crews dealing with serious accident scenarios. Studies in the behavioral literature [6] have documented the importance of two key features of simulator based training in speeding up the learning process. These are (1) the full control by the instructor of the stress level and (2) off-normal events and the capability for play-back in real, accelerated, and slowed time rates. However, while there is a very widespread belief in the effectiveness of simulator based training, studies which document and quantify this effectiveness in real emergency conditions have not yet been carried out due to the very limited data available under controlled conditions of crew response to an actual emergency without simulator training and with simulator training. Also, the importance of a high degree of physical fidelity is debated in the behavioral literature. While high fidelity has been found to be helpful for basic training of new operators by easing the transition from simulator to the real power plant, it seems to be less

important for highly experienced operators taking refresher training in emergency procedures.

UK Royal Navy Experience [7]

The United Kingdom Royal Navy has used simulators in training its nuclear power personnel assigned to the nuclear submarine fleet since 1971. Shown in Table 3 is a summary of the UK nuclear fleet for 1992 as reported in Janes 1993-94 with an estimate of the active crew requirements.

Table 3. 1992 UK Nuclear Ships

Type	Class	Active	First Active	Power (khp)	Displacement (tons)	Speed (kn)	Crew
SSBN	Vanguard	(4)	1993	27 (PWR2)	16,000	25	135
SSBN	Resolution	3	1967	15 (PWR1)	8,500	25	143
SSN	Trafalger	7+(6)	1983	15/27 (PWR1/2)	5,400	32	97
SSN	Swiftsure	5	1974	15 (PWR1)	4,900	30	116
SSN	Valiant	1	1966	15 (PWR1)	4,800	28	116
Total Active: 16 with 10 under construction or 15% U.S. submarine fleet							
Total Active Crew: Approximately 2,233 or about 12% U.S. active crew							

Engineering watch officers receive annual requalification training of 18-24 hours on a simulator split into 2 sessions about 6 months apart. As in the civilian nuclear power industry, this requalification training concentrates individual and crew responses to abnormal and accident situations. At the present time there are seven simulators in use provided by Marconi Simulation for four classes of nuclear submarines. These simulators provide exact fidelity to the submarine control room, audio effects for machinery operations and instructor-induced steam leaks, full record and play-back in real and non-real time, and linkage to maneuvering room simulators for combined exercises.

With over 20 years of simulator training experience for nuclear submarine personnel, the Royal Navy has been able to draw some important conclusions regarding the utility and effectiveness of using simulators. These are briefly summarized below.

Like the U.S. Navy, the Royal Navy has used land-based prototypes for training their nuclear power personnel. This practice was abandoned more than 10 years ago because safety restrictions made proper training of crews under emergency conditions impossible.

The highly controlled environment of the simulator has allowed the Royal Navy to compile an accurate data base of quantitative assessments of Engineering Watch officers' performance under reactor emergency conditions. This has been useful not only within the Royal Navy but as a data base for assessing human crew reliability for fault risk analysis for the general nuclear industry.

Analysis of crew errors under emergency conditions has led to ergonomic redesign of control panel instrumentation in UK submarines to reduce crew errors and enhance overall safety. A similar result occurred for the U.S. civilian nuclear power industry with the initial wave of plant-specific simulators purchase in the mid-1980s[8].

The Royal Navy has found there is an important synergy which develops between simulator experience and reactor operating experience. The models used for reactor analysis are better validated as the operating experience data base is expanded. A Rolls-Royce designed "black box" is installed on all UK nuclear submarines to document abnormal operating events as they occur. The loop is then closed by validating the simulator models against these events and incorporating them into the training program.

The Engineering Watch officer performance data base has found that months of routine operation at sea dulls crew emergency response performance. Consideration is being given to installation of ship-based simulators to reduce shore-based recertification time and to maintain a high level of training.

French Navy Experience [9]

The French nuclear navy is relatively small compared with the U.S. nuclear fleet but their experience seems to parallel that of the UK. The first reactor simulator was installed in 1971 for the SSBN Redoutable class and was decommissioned in 1985. In 1985 five simulators were installed for the SSBN L'Inflexible class at the Roland Morillot Training Center in Brest. In addition to the nuclear propulsion control station, these high-fidelity simulators also simulate the diving control with one axis mobility, navigation platform, tactical platform, and deterrent weapons platform. Simulator systems for the next generation of SSBNs are now under development. Shown in Table 4 is a summary of the French nuclear fleet for 1992 as reported in Janes 1993-94 with an estimate of the active crew requirements supported by their training facilities.

Table 4. 1992 France Nuclear Ships

Type	Class	Active	First Active	Power (khp)	Displace-ment (tons)	Speed (kn)	Crew
SSBN	Le Triomphant	(2)	-1995	41.5	14,120	25	111
SSBN	L'Inflexible	5	1973	16	8,920	25	114
SSN	Rubis	5+(1)	1983	9.5	2,670	25	70
Total Active: 10 with 3 under construction or 9% U.S. submarine fleet							
Total Active Crew: Approximately 1,490 or about 8% U.S. active crew							

INTRODUCTION OF AUTOMATION IN U.S. NUCLEAR NAVAL TRAINING

Given the very substantial advances in training simulation capability for nuclear power systems and more than 20 years of experience in both the civilian nuclear power industry and the UK and French nuclear navies, has the U.S. nuclear navy modified its historic opposition to these systems in training? The answer appears to be yes, although a quantitative assessment of efforts in this area was not possible due to a refusal by the U.S. Navy Division of Naval Propulsion to provide any information at all on this subject officially. However, what has been learned is that some initial steps have been taken by the U.S. Navy. These include the use of two retired SSBNs (Webster and Sam Rayburn) with reactor and missile tubes removed as training platforms at the submarine training center in Charleston, SC. In 1989-90 CAE-Link developed the first partial simulator for S5W plant for the Charleston training hulls. However, this was not a full engineering simulator since it covered only a part of the nuclear plant controls. In fact, the Navy does not refer to it as a simulator, but rather as a "training board." At the present time at least one of the major simulator manufacturers in the U.S. is under contract to pursue simulator related work, but the details of this work are not publicly available.

The conclusions we can draw from the limited information available, is that the U.S. Navy is moving tentatively into this area. However, it does so after 20 years of opposition to using automated systems not only in training but in the automation of systems in the engineering space of nuclear submarines. As a consequence, the experience base needed for implementing state-of-the-art control systems in the next generation submarines has not been developed.

ENGINEERING SPACE AUTOMATION IN SUBMARINES— CENTURION DESIGN

The long-term effects of the recently relaxed policy opposing automated systems for training and for installation in the engineering space of U.S. nuclear submarines is clearly illustrated by the current debate over the design of the next generation attack submarine, the Centurion. Navy Assistant Secretary Gerald Cann has been quoted as saying that inclusion of a fully automated engineering space in the Centurion design is ruled out as premature: "I don't think that's in the cards yet. In order to do something like that, we'd have to have a serious prototyping effort and get lots of experience..." [10]. The practical consequence of this is that the overall power density of our nuclear submarines is low and a larger crew size is required for operation. Examination of data for the UK Royal Navy in Table 3 clearly shows that the newest submarines have increased power and reduced crew size, while U.S. submarines have maintained a crew of 133 through the Los Angeles and Seawolf class attack submarines.

With low power density and larger crew size, the only way to maintain acceptable performance in attack submarines (i.e., top speed) is to build a much larger and costly ship. To try to quantify this effect, we can develop a very simple model for submarines of similar form factor (length to diameter ratio) where we relate the top speed of a submarine to the cube root of the ratio of the propulsion power to the submarine surface area. For fixed form factor, the surface area is proportional to the $2/3$ power of the submarine's submerged displacement, so that:

$$\text{Speed} = C \sqrt[3]{\frac{\text{Power}}{[\text{Displacement}]^{2/3}}}$$

With the units of power in horsepower and submerged displacement in tons we can evaluate the constant C for a wide variety of submarines as shown in Table 5 below using only published data from Janes 1993–94.

**Table 5. Evaluation of Constant, C,
Relating Speed to Power and Size**

Country	Type	Class	Power (khp)	Displacement (tons)	Speed (kn)	C
U.S.	SSN	Seawolf	52	9,137	35	7.12
U.S.	SSN	Los Angeles	35	6,927	32	7.00
U.S.	SSN	Sturgeon	15	4,960	30	8.08
U.S.	SSBN	Ben Franklin	15	8,250	25	7.54
U.S.	SSBN	Ohio	60	18,770	?20+	?4.55+
UK	SSN	Trafalgar	27.5	5,900	32	7.31
UK	SSN	Swiftsure	15	4,900	30	8.05
UK	SSN	Valiant	15	4,800	28	7.48
UK	SSBN	Vanguard	27.5	16,000	25	7.14
UK	SSBN	Resolution	15	8,500	25	7.59
France	SSBN	L'Inflexible	16	8,920	25	7.49
France	SSN	Rubis	9.5	2,670	25	6.82

We see from Table 5 that our simple model works remarkably well over a very wide range of nuclear submarines ranging from the 2,670-ton French Rubis SSN to the 16,000-ton UK Vanguard SSBN, with C in the range from about 7 to 8. The notable exception is the Ohio class SSBN which has a somewhat larger form factor than the others and is only listed in Janes as capable of greater than 20 knots. One thing to also note is that there is a trend towards $C \sim 8$ for older submarines which are presumably less quiet and values close to $C \sim 7$ for the newer generation of quieter submarines. This reduction in the value of C presumably reflects a drag penalty for employing improved noise signature reduction techniques.

Let's consider a smaller and less costly approach to the Centurion design in the 5000 ton class similar to the Sturgeon class SSNs using a 15 khp S5W class nuclear plant. Assuming a reduction in C from 8 to 7 or possibly less to account for the drag penalty of state-of-the-art quieting technology, the top speed would be reduced to 26 knots for $C=7$ and 24 knots if C were as low as 6.5. This is quite a bit slower than the 30 knot class SSNs currently deployed. The only way to bring the speed up to 30 knots with $C \leq 7$ is to increase size to accommodate more power moving toward Seawolf size submarines. However, if a fully automated engineering space could increase power density and reduce crew requirements to below 100, allowing a 20 khp plant to be installed in a 5,000-ton submarine, then the top speed at $C=7$ would be increased to 29 knots and open up the possibility of a viable design in this much smaller and less costly size class.

DISCUSSION AND RECOMMENDATIONS

The technical and strategic context faced by the U.S. nuclear navy today as it considers what policy it should pursue regarding the use of automated systems both for training and submarine design is very much different from what existed when Rickover established a ban on using this technology. The key elements affecting the U.S. nuclear navy are as follows:

- The Navy has been discussing plans to decommission up to 50 nuclear ships over the next 4 years.
- New submarine production is being justified primarily to maintain the nuclear submarine industrial base, not to match a growing foreign threat.
- Automated systems for control and training are highly developed in the commercial nuclear industry and in foreign nuclear navies—but not in the U.S. nuclear navy.
- Cost considerations will likely force the closure of most land-based prototype nuclear reactors in the near future.

Two compensating trends will affect the training needs of the U.S. nuclear navy in the next few years. The reduction in fleet size will reduce training needs, but the loss of land-based prototypes will also reduce the number of training facilities. It is interesting to compare the number of full-function reactor simulators needed to give our submarine crews the same level of training as received by the UK and French submarine crews both of which provide 1 simulator for roughly each 300 active crew members. Given the size of the U.S. nuclear fleet in 1992, this would require about 57 simulators to achieve the same level of training! Even if we consider the projected reductions in the nuclear fleet by 1997, this would still require in the neighborhood of 30 simulators for the S6G and S8G reactors.

In the interests of both maintaining the highest level of safety, as well as anticipating the loss of prototypes in the training program, the U.S. Navy should move aggressively to establish an integrated simulator-based training program. If the power of the recently introduced generation of processors allows installation of ship-based simulators for maintaining a high level of emergency response readiness, the need for the large numbers of shore-based systems estimated above would be substantially reduced with significant cost savings possible. Since studies have shown a significant reduction in the timely crew response to emergency situations after months of largely routine operation, such a ship-based simulator would be used as a training device to ensure a high level of crew readiness for emergency transient response. Additionally, the computing power

available in state-of-the-art simulators is now sufficient to provide an analysis tool for the guidance of engineering officers during the evolution of an actual serious reactor casualty.

The introduction of high fidelity simulators of the reactor and engineering control panels in nuclear navy training centers should also be exploited as a laboratory to develop improved panel layout for minimization of operator errors and improved crew coordination based on quantitatively documented data obtained during actual crew training, and to test various levels of automation of the engineering spaces in submarines and assess implications on crew requirements. In addition, it may be possible to connect the reactor control simulator stations to the existing command and fire control simulators to allow full scale exercises integrating engineering and tactical response under normal and various levels of casualty conditions.

Finally, in view of the fact that there is no immediate threat justification for embarking on a large production run of a new attack submarine, this would seem to be a very opportune time to build one or more prototypes to bring the automation level in the engineering systems up to the current state of the art, leading to a smaller and less costly nuclear attack submarine design for large scale production. If prototype production can sustain the industrial base in the short term, it would be more cost-effective than committing now to the large scale production of an expensive replacement for the Los Angeles class SSNs incorporating obsolete technology for the nuclear propulsion systems.

REFERENCES

1. Norman Polmar and Thomas B. Allen, *Rickover, Controversy and Genius, A Biography*, Simon & Schuster, New York, 1982.
2. Patrick Tyler, *Running Critical*, Harper & Row, New York, 1986.
3. Francis Duncan, *Rickover and the Nuclear Navy*, Naval Institute Press, Annapolis, 1990.
4. Testimony of Adm. H. G. Rickover to Joint Economic Committee, *Economics of Defense Policy*, Part 1, 97th Congress, 2nd Session, Washington, DC, 1982, p. 97.
5. *Navy Times*, 3 March 1965.
6. R.B. Stammers, "Instructional Psychology and the Design of Training Simulators," *Simulation for Nuclear Reactor Technology*, (Cambridge University Press, Cambridge, 1985, D. G. Walton, editor) pp. 161-176.
7. A.J.H. Burbridge, "Use of Simulators for the Continuation Training of Nuclear Propulsion Plant Operators in the Royal Navy," *Simulation for Nuclear Reactor Technology*, (Cambridge University Press, Cambridge, 1985, D.G. Walton, editor) pp. 195-215.
8. R.A. Felker, S3 Technologies Company, private communication.
9. G. Valason, "Simulateurs pour Sous-Marins," *Defense National*, June 1990, pp. 183-186.
10. *Navy News & Undersea Technology*, March 2, 1992, p. 8.

**APPENDIX A
MEMBERS (1992-1993)**

Appendix A

MEMBERS (1992-1993)

PETER CHEN
Department of Chemistry, Harvard University

Ph.D. (Chemistry), Yale University, 1987

RESEARCH INTERESTS:
Laser spectroscopy and kinetics of organic and inorganic reactive intermediates, photoelectron and photo ionization mass spectroscopic thermochemical measurements, and applications of supersonic jet expansions to physical organic chemistry

WILLIAM J. DALLY
Department of Electrical Engineering and Computer Science,
Massachusetts Institute of Technology

Ph.D. (Computer Science), California Institute of Technology, 1986

RESEARCH INTEREST:
Concurrent architecture's, interconnection networks, VLSI architecture, special purpose processors, architecture experiments, computer aided design

MARK E. DAVIS
Department of Chemical Engineering, California Institute of Technology

Ph.D. (Chemical Engineering), University of Kentucky, 1981

RESEARCH INTERESTS:
Synthesis of zeolites and molecular sieves; catalysis involving zeolites and molecular sieves; novel heterogeneous catalysts; novel catalytic reactor configurations

S. JAMES GATES, JR.
Department of Physics and Astronomy, University of Maryland - and -
Department of Physics and Astronomy, Howard University

Ph.D. (Physics), Massachusetts Institute of Technology, 1977

RESEARCH INTERESTS:
Quantum field theories concentrating in special category of supersymmetric theories; investigation into gauge field theories, gravitation and general relativity with applications to the physics of elementary particles

MEMBERS (continued)

NANCY M. HAEGEL

Department of Materials Science and Engineering, University of California, Los Angeles
Department of Physics, Fairfield University

Ph.D. (Materials Science), University of California, Berkeley, 1985

RESEARCH INTERESTS:

Optical and electrical characterization of semiconductors, electrical transport in high resistivity semiconductors, infrared detectors and photoconductors

THOMAS C. HALSEY

The James Franck Institute and Department of Physics, University of Chicago

Ph.D. (Physics), Harvard University, 1984

RESEARCH INTERESTS:

Dynamics and statistical mechanics of Josephson junction arrays. Diffusive growth and pattern formation; multifractality, hierarchical models of the active surface, generalized aggregation models. Physical problems in electrochemistry: the double layer impedance, pattern formation in electrodeposition. Dynamical systems: scaling structure of strange attractors. Dynamics of crack propagation and fracture mechanics. Dynamics and statistical mechanics of electro-rheological fluids.

ROBERT A. HUMMEL

Department of Computer Science, Courant Institute of Mathematical Sciences,
New York University

Ph.D. (Mathematics), University of Minnesota

RESEARCH INTEREST:

Computer vision, model matching, autonomous target recognition, parallelism, uncertainty reasoning, modeling and simulation, variational methods

KEVIN K. LEHMANN

Frick Chemical Laboratories and Department of Chemistry, Princeton University

Ph.D. (Chemical Physics), Harvard University, 1983

RESEARCH INTERESTS:

High resolution laser spectroscopy: Development of double resonance techniques for the study of excited vibrational and electronic states of polyatomic molecules; determination of the magnitude of intermode coupling constants or intramolecular relaxation rates; study the consequences of the onset of chaos for quantum mechanical systems and look for "quantum chaos" in molecular spectra.

MEMBERS (continued)

DAVID L. McDOWELL

School of Mechanical Engineering, Georgia Institute of Technology

Ph.D. (Mechanical Engineering), University of Illinois at Urbana-Champaign, 1983

RESEARCH INTERESTS:

Inelastic deformation and fracture of engineering and advanced materials, high temperature deformation and damage processes, damage mechanics and finite deformation inelasticity

ANNE B. MYERS

Department of Chemistry, University of Rochester

Ph.D. (Chemistry), University of California, Berkeley, 1984

RESEARCH INTERESTS:

Resonance Raman spectroscopy, molecular dynamics in gas and solution phases, four-wave mixing spectroscopies, picosecond spectroscopies

GERALD A. NAVRATIL

Department of Applied Physics, Columbia University

Ph.D. (Plasma Physics), University of Wisconsin, Madison, 1976

RESEARCH INTERESTS:

Experimental plasma physics, controlled thermonuclear fusion research, development of plasma diagnostic techniques

ROBERT A. PASCAL, JR.

Frick Chemical Laboratory and Department of Chemistry, Princeton University

Ph.D. (Biochemistry), Rice University, 1980

RESEARCH INTERESTS:

Enzymatic reaction mechanisms; bioorganic chemistry; biosynthesis of naturally-occurring compounds, biochemistry of parasites; physical organic chemistry; synthesis of structurally unusual organic compounds

DENNIS L. POLLA

Department of Electrical Engineering, University of Minnesota

Ph.D. (Electrical Engineering), University of California, Berkeley, 1985

RESEARCH INTERESTS:

Silicon based micromechanics (microsensors and microactuators), HgCdTe infrared detectors, and VLSI microelectronics

MEMBERS (continued)

PETER W. VOORHEES

Department of Materials Science and Engineering, Northwestern University

Ph.D. (Materials Engineering), Rensselaer Polytechnic Institute, 1982

RESEARCH INTERESTS:

Phased transformations, crystal growth, elastic stress effects during phased tranformations,
Ostwald ripening

APPENDIX B
MENTORS AND ADVISORS (1992-1993)

Appendix B MENTORS AND ADVISORS (1992-1993)

DANIEL ALPERT	Director, Program in Science, Technology & Society, University of Illinois
R. STEPHEN BERRY	Professor of Chemistry, University of Chicago
SOLOMON J. BUCHSBAUM*	Senior Vice President, Technology Systems, AT&T Bell Laboratories
CURTIS CALLAN	Professor of Physics, Princeton University Chairman of JASON
RUTH DAVIS	President, The Pymatuning Group, Inc.
RUSSELL E. DOUGHERTY	General, U.S. Air Force (Retired) Private Consultant
ALEXANDER H. FLAX	Private Consultant
ANDREW J. GOODPASTER	General, U.S. Army (Retired) Private Consultant
PAUL F. GORMAN	General, U.S. Army (Retired) President, Cardinal Point, Inc.
ISAAC C. KIDD, Jr.	Admiral, U.S. Navy (Retired) Private Consultant
STEVEN E. KOONIN	Professor of Theoretical Physics, California Institute of Technology
MARTHA KREBS	Associate Director, Planning & Development, Lawrence Berkeley Laboratory
STANFORD S. PENNER	Director, Center for Energy and Combustion Research, University of California, San Diego
DAVID PINES	Professor of Physics and Electrical Engineering, University of Illinois

* Deceased March 1993

MENTORS AND ADVISORS (continued)

ROBERT E. ROBERTS

Vice President-Research,
Institute for Defense Analyses

WILLIAM Y. SMITH

General, U.S. Air Force (Retired)
Private Consultant

HARRY D. TRAIN, II

Admiral, U.S. Navy (Retired)
Private Consultant

LARRY D. WELCH

General, U.S. Air Force (Retired)
President, Institute for Defense Analyses

HERBERT YORK

Director Emeritus, The Institute on Global
Conflict and Cooperation,
University of California, San Diego

APPENDIX C
DSSG MANAGEMENT TEAM (1992-1993)

Appendix C
DSSG MANAGEMENT TEAM (1992-1993)

CHAIRMAN

General W. Y. Smith
USAF (Retired)

PROGRAM DIRECTOR

Dr. Julian C. Nall

ASSOCIATE PROGRAM DIRECTOR

Dr. William J. Hurley

PROGRAM ADMINISTRATOR

Ms. Nancy P. Licato

APPENDIX D
PROGRAM PLAN (1992-1993)

Appendix D

PROGRAM PLAN FOR THE

DEFENSE SCIENCE STUDY GROUP (1992-1993)

PURPOSE

To foster a long-term interest in national security issues among DSSG members, and to facilitate their continued involvement with such issues.

GENERAL

- Program of education and study related to national security
- Flexible to accommodate the changing times, guidance of the mentors, and special interests of the members
- Two years
- Approximately 20 days per year
- Four sessions each year

FIRST YEAR OF PROGRAM

SESSION 1: Introduction to National Security Issues

General

- Three days (two days at IDA, one day for local visits)
- Probably will begin on Wednesday during late February or early March
- All mentors invited

Briefings

- Introduction to the DSSG program
- DoD--especially DIA, ARPA, R&D programs of the military services, etc.
- The Intelligence Community
- Policy Making
- IDA
- What to expect--by an alumnus or alumna
- Senior academic who contributes to national security

Visits

- DoD--such as the National Military Command Center, R&D facilities, program offices, etc.

Interaction with mentors

- General discussion with mentors

SESSION 2: Air Force, Marines, and Industry**General**

- Seven weekdays plus weekend
- Probably will begin on next to the last Monday in June
- With some mentors

Briefing

- Introduction to the Air Force and Marines

Visits

- Air Force facilities
 - Such as STRATCOM, NORAD, Red Flag, etc.
- Marine facilities
 - Such as Camp Pendleton or Marine Air-Ground Training Center (or Camp Lejeune later)
- West Coast industrial facilities
 - Manufacturing technologies
 - High-tech R&D

SESSION 3: Army, Navy, and Industry**Briefing**

- Introduction to the Army and Navy

General

- Seven weekdays plus weekend
- Probably will begin the first Monday in August
- With some mentors

Visits

- Army facilities
 - Such as Fort Bragg
- Navy related facilities
 - Such as Electric Boat, Norfolk, and Kings Bay
 - Ship at sea
- East Coast industrial facilities
 - Manufacturing technologies
 - High-tech R&D

SESSION 4: Study Topics

General

- Three days at IDA
- Probably will begin on the third Monday in November
- All mentors invited

Study Topics

- Discussion with mentors about possible "initial inquiries" or "think pieces" on development of new ideas of potential importance to national security and of high interest to the member
- Selected from one or more topics suggested prior to session by each member or team of members
- Where appropriate, one or more mentors will help provide general guidance to each study

Discussion by members

- Experiences during first year of course

Briefings

- A few by senior officials

SECOND YEAR OF PROGRAM

SESSION 5: The Intelligence Community

General

- Three days in Washington area
- Probably will begin on the second Wednesday in March
- With some mentors

Visits

- Central Intelligence Agency
- Defense Intelligence Agency
- National Security Agency

SESSION 6: National Security Issues, Study Topics, and Industry

General

- Seven weekdays plus weekend
 - Primarily in Washington
 - A few days on short tour
- Probably will begin on next to the last Monday in June
- Some mentors may attend

Briefings

- One or two by senior officials from the national security community

Study Topics

- About four days plus weekend for individual and group work on "initial inquiries" or "think pieces"
- Individual and group discussions with persons in the Washington area who can assist in work on study topics

Visits

- About three days at industrial facilities
 - Manufacturing technology
 - High-tech R&D

SESSION 7: Preparation of "Initial Inquiries" or "Think Pieces"

General

- Seven weekdays plus weekend
- Probably will begin on the first Monday in August
- At a National Laboratory and visit to the National Training Center
- Some mentors who have been helping with studies may attend part of the session

Studies

- Completion of draft "initial inquiries" or "think pieces"

Briefings and Tours

- Selected areas of interest at National Laboratory

Visit

- National Training Center

SESSION 8: Final Session

General

- Three days at IDA
- Probably will begin on third Sunday in November
- All mentors invited for second and third days

Preparation

- Final preparation for presentations--briefing aids, etc.

Presentations

- "Initial inquiries" or "think pieces" presented by members to mentors, representatives from ARPA and IDA, and to fellow members.

Briefings

- One or two senior officials from DoD or non-DoD agencies

Graduation and reception

APPENDIX E
ALUMNI (1992-1993)

Appendix E ALUMNI (1992-1993)

STEPHEN P. BOYD	Department of Electrical Engineering, Stanford University
RUSSEL E. CAFLISCH	Department of Mathematics, University of California, Los Angeles
STEPHEN A. CAMPBELL	Department of Electrical Engineering, University of Minnesota
STEVEN K. CASE	Department of Electrical Engineering, University of Minnesota and CyberOptics Corporation
VICKI L. CHANDLER	Institute of Molecular Biology, University of Oregon
SUSAN N. COPPERSMITH	Optical Materials and Interface Research Department, AT&T Bell Laboratories
WERNER J.A. DAHM	Department of Aerospace Engineering, University of Michigan
ROBERT H. DAVIS	Department of Chemical Engineering, University of Colorado
KATHERINE T. FABER	Department of Materials Science and Engineering, Northwestern University
JOSEPH S. FRANCISCO	Department of Chemistry, Wayne State University
STEVEN M. GEORGE	Department of Chemistry and Biochemistry, University of Colorado
BRUCE HAJEK	Department of Electrical and Computer Engineering, University of Illinois
JAMES M. HOWE	Department of Materials Science and Engineering, University of Virginia
DEBORAH A. JOSEPH	Department of Computer Science, University of Wisconsin
RANDY H. KATZ	Department of Electrical Engineering and Computer Science, University of California, Berkeley

DSSG ALUMNI (continued)

STEVEN E. KOONIN	Division of Physics, Mathematics, and Astronomy, California Institute of Technology
FREDERICK K. LAMB	Department of Physics, University of Illinois
NATHAN S. LEWIS	Department of Chemistry, California Institute of Technology
PHILIP S. MARCUS	Department of Mechanical Engineering, University of California, Berkeley
DANIEL M. NOSENCHUCK	Department of Mechanical and Aerospace Engineering, Princeton University
ANTHONY T. PATERA	Department of Mechanical Engineering, Massachusetts Institute of Technology
THOMAS A. PRINCE	Division of Physics, Mathematics, and Astronomy, California Institute of Technology
THOMAS F. ROSENBAUM	Department of Physics, University of Chicago
STEPHEN W. SEMMES	Department of Mathematics, Rice University
STEVEN J. SIBENER	Department of Chemistry, University of Chicago
THEODORE A. SLAMAN	Department of Mathematics, University of Chicago
DANIEL L. STEIN	Department of Physics, University of Arizona
WARREN S. WARREN	Department of Chemistry, Princeton University
ROBERT L. WHETTEN	School of Physics, Georgia Institute of Technology
R. STANLEY WILLIAMS	Department of Chemistry and Biochemistry, University of California, Los Angeles
W. HUGH WOODIN	Department of Mathematics, University of California, Berkeley

APPENDIX F
SOME DEFENSE RELATED ACTIVITIES OF DSSG
ALUMNI (1992-1993)

Appendix F
SOME DEFENSE RELATED ACTIVITIES OF DSSG
ALUMNI (1992-1993)

RUSSEL E. CAFLISCH

Research Grant on Fluid Dynamics, Kinetic Theory and Monte Carlo Methods for Air Force Office of Scientific Research

Research Grant on High Reynolds Number Fluid Flows for ARPA

Research Grant on Nonlinear and Stochastic Numerical Methods for Army Research Office

Participant, Future Technology Panel, Naval Studies Board Review of Future Aircraft Carrier Design

STEPHEN A. CAMPBELL

Participant, Defense Science Board (DSB) Task Force on Weapon Development and Production Technology, 1991

STEVEN K. CASE

National Academy of Engineering, Task Force on Design of the Next Generation Aircraft Carriers

NASA, Laser Based Sensors for the Inspection of Solid Rocket Booster O-Ring Seals

U.S. Navy, Use of Laser Sensors for Measuring the Effectiveness of Explosives

Army Research Office, Laser Measurement of Ammunition Manufacturing Quality

WERNER J. A. DAHM

Participant, Defense Science Board (DSB) Summer Study, Tactical Air Warfare Task Force, Operational Concepts and Force Configuration Panel, 1993

Participant, IDA Central Research Project and co-author, IDA Paper P-2783, AS-IS (Active Safing & Isolation System): A Satellite-Based Remote Continuing Authorization Concept with Application to Control of Naval Strategic Nuclear Missiles and Tactical Weapons," June 1992

WILLIAM J. DALLY

Member of JASON, April 1993-present

ROBERT H. DAVIS

Attended ONR Disarmament Symposium, February 1993

Reviewed ONR Program on Advanced Propellant Processing, June 1993

Wrote proposal to U.S. Army on Advanced Membrane Technology for Water Purification, November 1993

STEVEN M. GEORGE

Member of National Research Council Committee: National Materials Advisory Board, study of "Next Generation Currency Design," charged with determining ways to prevent counterfeiting, 1992 and 1993

Member of National Research Council Committee: Board of Assessment of NIST Activities and on the Chemical Science and Technology subpanel, 1993-1996

BRUCE HAJEK

Participant, Communication Networks, Joint Services Electronics Program

DEBORAH A. JOSEPH

Member, Computer Science and Telecommunications Board, National Research Council, 1993-present

RANDY H. KATZ

Co-chairman, ARPA ISAT (Information Science and Technology), Summer Study on High Performance Memory Systems, August 1991

Participant, ARPA ISAT Summer Study on Instant Infrastructure, August 1992

Co-chair, White House Technology Task Force, 1993

Participant, National Performance Review, "Reengineering Government through Information Technology" Team 1993

Co-chairman, ARPA ISAT Summer Study on National Information Infrastructure Services, August 1993

Co-chairman, High Performance Computing, Communications, and Information Technology (HPCCIT) Program Information Infrastructure Technology and Applications (IITA) Task Group, 1993

Program Manager, "National-scale Information Enterprise Program," Computer Systems Technology Office, ARPA, 1993 and 1994 [on sabbatical from University of California, Berkeley].

STEVEN E. KOONIN

Member, JASON, 1988-present

Sample of study topics: Brilliant Pebbles, technology for arms control verification, weapon-effects simulators, tagging explosives for countering terrorists

Member, Defense Science Board, 1991-present

Mentor, Defense Science Study Group, 1992-present

DoE Hydrotest Program Assessment, 1991-1992

DoE Inertial Confinement Fusion Advisory Committee, 1992-present

DoE Task Force on Energy Research Priorities, 1991-1992

Participant, Defense Science Board Task Force on Strategic Sensors, 1990

Chairman of a large review of Intertial Confinement Fusion for NRC. This study was performed (early 1990) at the request of the Secretary of Energy as mandated by the House Armed Services Committee (HASC). Briefed both the HASC and Secretary Wallins.

FREDERICK K. LAMB

Since 1986 have been working on verification of underground nuclear tests that began due to briefings received on arms control while a member of DSSG for ARPA, Congressional Office of Technology Assessment, DOE, ACDA, and others

Worked with Congressional OTA to examine methods to measure the yield of underground nuclear tests; an ARPA/NMRO and Air Force Geophysics Laboratory funded program. This program examined hydrodynamic methods for measuring the yield, and in particular the CORTEX method.

Worked on ARPA/NMRO funded program at IDA on assessing methods for monitoring underground nuclear tests and authored, IDA Document D-1363, "Effects of Nuclear Devices and Device Canisters on the Accuracy of Hydrodynamic Yield Estimates for Threshold Test Ban Treaty and Peaceful Nuclear Explosions Treaty Verification (U)," Secret, October 1991

Participant in DOE Symposium on Explosion Source Phenomenology , Mach 1989

Served on the Red Team for Arms Control Disarmament Agency/National Security Council review of the Peaceful Nuclear Explosions Treaty and Threshold Test Ban Treaty Protocols, summer 1989

Serves on the Executive Committee of the University of Illinois Program on Arms Control, Disarmament, and International Society

NATHAN S. LEWIS

Member, JASON, 1990-present

Led studies on Nonproliferation and on Underground Sensors, and contributed to numerous other studies on topics, including Hanford waste cleanup, military applications for biomedical technology, and processes for verification of weapons dismantlement

Member of Visiting Committee, Brookhaven National Laboratory, Department of Applied Science

PHILIP S. MARCUS

Member, Physical Sciences Advisory Committee (PSAC), Lawrence Livermore National Laboratory, DoE. (Involves a semi-annual review of programs under the physics directorate)

Invited to join JASON in 1988

DANIEL M. NOSENCHUCK

Involved in Program Reviews of Submarine Technology Program at ARPA/Maritime Systems Technology Office

Co-author of IDA Central Research Project, IDA Paper P-2783, "AS-IS (Active Safing & Isolation System): A Satellite-Based Remote Continuing Authorization Concept with Application to Control of Naval Strategic Nuclear Missiles and Tactical Weapons," June 1992

Participant, Defense Science Board (DSB) Task Force on Ballistic Missile Defense, 1991

ANTHONY T. PATERA

Member of Defense Sciences Research Council, ARPA (formerly Materials Research Council)

THOMAS A. PRINCE

Member, JASON, 1991-present

THOMAS F. ROSENBAUM

Participant, Defense Science Board Task Force on Improving Testing and Evaluation Effectiveness, 1989

THEODORE A. SLAMAN

Participant, Defense Science Board (DSB) Task Force on Defense Technology Strategies, 1991

DANIEL L. STEIN

Participant, Defense Science Board Task Force on National Security Space Launch Strategy, 1989

WARREN S. WARREN

Worked for Defense Sciences Office, ARPA, on Microwave and Optical Pulse Shaping

R. STANLEY WILLIAMS

Invited speaker at the Defense Sciences Research Council Workshop on "Statistical Limits of Ultra-Small Devices," LaJolla, CA, July 12-13, 1993, ARPA

Research contract from ONR Chemistry Division: "Laser Ablation as a New Synthetic Route for Metastable Materials."

Research Contract from ONR Electronics Division: "Atomic Layer Epitaxy with Supersonic Molecular Beams: Diamond and Group IV Compounds"

W. HUGH WOODIN

Consultant to IDA Center for Communications Research (CCR)

APPENDIX G
GLOSSARY

Appendix G

GLOSSARY

ACDA	Arms Control and Disarmament Agency
AEC	Atomic Energy Commission
AFB	Air Force Base
AFQT	Armed Forces Qualifying Test
AIT	Advanced Individual Training
ALE	arbitrary Lagrangian-Eulerian
ANEC	American Nuclear Energy Council
ANFO	ammonium nitrate/fuel oil
APC	armored personnel carrier
ARPA	Advanced Research Projects Agency
ASVAB	Armed Services Vocational Aptitude Battery
ATA	alimentary toxic aleukia
ATR	automatic target recognition
BWR	boiling water reactor
CE	chemical energy
CID	combat identification
cm	centimeter
DCI	Director of Central Intelligence
DDR&E	Director, Defense Research and Engineering
DIA	Defense Intelligence Agency
DOE	Department of Energy
DSP	Defense Support Program
DSSG	Defense Science Study Group
DU	depleted uranium
EFP	explosively formed projectile
EM	electromagnetic
ERDA	Energy Research and Development Agency
ESR	electroslag remelting
FAA	Federal Aviation Administration
FFRDC	Federally Funded Research and Development Center
FLIR	forward looking infrared
FLIT	Functional Literacy (U.S. Army program)
FSAR	Final Safety Analysis Report
GPR	ground penetrating radar
GPS	Global Positioning System
HE	high explosive
HEL	Hugoniot Elastic Limit

IDA	Institute for Defense Analyses
IFF	identification friend or foe
IR	infrared
JOBS	Job Oriented Basic Skills
JSEP	Job Skills Education Program
JSTARS	Joint Surveillance, Target, Attack Radar System
kbits/s	thousand bits per second
KE	kinetic energy
kHz	kilo hertz
km	kilometer
LCAC	landing craft air cushion
LPI	low probability of intercept
m	meter
MajGen	Major General (Air Force, Marine Corps)
MEMS	microelectromechanical system
MG	Major General (Army)
MILSTAR	Military Strategic-Tactical and Relay Satellite System
mm	millimeter
MMW	millimeter wave
ms	millisecond
MTBF	mean time between failures
NAS	National Academy of Sciences
NDE	nondestructive evaluation
NDT	nondestructive testing
NEC	National Electromagnetic Code
NMAB	National Materials Advisory Board
NORAD	North American Air Defense
NRC	National Research Council
NRC	Nuclear Regulatory Commission
NRTS	Nuclear Reactor Testing Station
NSA	National Security Agency
NWPA	Nuclear Waste Policy Act
ODS	Operation Desert Storm
OMB	Office of Management and Budget
OTA	Office of Technology Assessment
OTH	over the horizon
PAN	peroxyacetyl nitrate
PFNA	pulsed fast neutron activation
PIS	position information system
ppm	parts per million
PWR	pressurized water reactor
RAH	rolled homogeneous armor
ROE	Rules of Engagement

S/N	signal to noise
SA	situational awareness
SAW	surface acoustic wave
SIMNET	simulation network
SNR	signal to noise ratio
SpEcBar	spatially encoded barcode
TMI	Three Mile Island
TNA	thermal neutron activation
TRADOC	Training and Doctrine Command
USA	U.S. Army
USAF	U.S. Air Force
VIM	vacuum induction melting
XFSP	Experimental Functional Skills Program
μ s	microsecond

APPENDIX H
DISTRIBUTION LIST FOR IDA PAPER P-2949
Volume I

Appendix H
DISTRIBUTION LIST FOR IDA PAPER P-2949
Volume I

	<u>Number of Copies</u>
GOVERNMENT AGENCY	
Advanced Research Projects Agency 3701 N. Fairfax Drive Arlington, VA 22203-1714	
ATTN: Dr. Duane A. Adams, Deputy Director	1
Dr. Ira D. Skurnick, DSO/AS	1
DSSG MEMBERS OF CLASS III	
Professor Peter Chen Laboratorium fuer Organische Chemie ETH Zentrum Universitatstrasse 16 CH-8092 Zurich Switzerland	1
Professor William J. Dally Artificial Intelligence Laboratory Massachusetts Institute of Technology 545 Technology Square Cambridge, MA 02139	1
Professor Mark E. Davis Chemical Engineering 210-41 California Institute of Technology Pasadena, CA 91125	1
Professor S. James Gates, Jr. Department of Physics and Astronomy University of Maryland College Park, MD 20742-4111	1
Professor Nancy M. Haegel Department of Physics Fairfield University Fairfield, CT 06431	1
Professor Thomas C. Halsey Exxon Research and Engineering Route 22 East Annandale, NJ 08801	1
Professor Robert A. Hummel Courant Institute of Mathematical Sciences New York University 251 Mercer Street New York, NY 10012	1

DSSG Members of Class III (cont.)Number of Copies

Professor Kevin K. Lehmann Frick Chemical Laboratory Princeton University Washington Road Princeton, NJ 08544-1009	1
Professor David L. McDowell The George W. Woodruff School of Mechanical Engineering Georgia Institute of Technology Atlanta, GA 30332-0325	1
Professor Anne B. Myers Department of Chemistry University of Rochester Rochester, NY 14627	1
Professor Gerald A. Navratil Department of Applied Physics Columbia University 215 S.W. Mudd Building New York, NY 10027	1
Professor Robert A. Pascal, Jr. Frick Chemical Laboratory Princeton University Washington Road Princeton, NJ 08544-1009	1
Professor Dennis L. Polla Department of Electrical Engineering University of Minnesota 200 Union Street, SE Minneapolis, MN 55455	1
Professor Peter W. Voorhees Department of Materials Science and Engineering Northwestern University Evanston, IL 60208	1

OTHER ORGANIZATIONS

Institute for Defense Analyses 1801 N. Beauregard St. Alexandria, VA 22311-1772	25
Defense Technical Information Center Cameron Station Alexandria, VA 22314	2

REPORT DOCUMENTATION PAGE

*Form Approved
OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)			2. REPORT DATE December 1994		3. REPORT TYPE AND DATES COVERED Final	
4. TITLE AND SUBTITLE Summary Report of the Defense Science Study Group III, 1992-1993, Volume I (U)			5. FUNDING NUMBERS DASW-01-94-C-0054 A-103			
6. AUTHOR(S) N.P. Licato						
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Institute for Defense Analyses 1801 N. Beauregard Street Alexandria, VA 22311			8. PERFORMING ORGANIZATION REPORT NUMBER IDA Paper P-2949			
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Advanced Research Projects Agency 3701 N. Fairfax Drive Arlington, VA 22203-1714 Attn: Dr. Ira Skurnick, DSO/AS			10. SPONSORING/MONITORING AGENCY REPORT NUMBER FFRDC Programs 2001 N. Beauregard Street Alexandria, VA 22311			
11. SUPPLEMENTARY NOTES						
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution is unlimited; 24 May 1995.			12b. DISTRIBUTION CODE			
13. ABSTRACT (Maximum 200 words) The Defense Science Study Group (DSSG) is a program sponsored by the Advanced Research Projects Agency (ARPA) and managed by the Institute for Defense Analyses (IDA). The purpose of the program is to identify some of the most talented young scientists and engineers from academia and expose them to issues and operations related to national security. By strengthening ties between the scientists and engineers and the national security community it is hoped that the Government will be provided with a new source of technical advisors and informed critics.						
Individuals spend about 20 days per year for 2 years receiving briefings by distinguished speakers; visiting laboratories, intelligence organizations, and military, manufacturing, and industrial facilities; and conducting studies on defense-related topics. A group of mentors who have distinguished careers in defense, industry, or academia provide guidance to the program. This report summarizes the program's activities from 1992-1993.						
14. SUBJECT TERMS Defense Science Study Group, defense technology, defense-related academic research, defense science research, aircraft component fatigue monitoring, identification friend or foe (IFF), automatic target recognition (ATR), chemical and biological warfare (CBW), ballistic impact and blast waves			15. NUMBER OF PAGES 337		16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED		18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED		19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED		
20. LIMITATION OF ABSTRACT Same as report						